이 학 박 사 학 위 논 문

Asymptotic Analysis of Machine Learning Algorithms

기계 학습 알고리즘의 점근 성질 연구

2020년 2월

서울대학교 대학원

통계학과

온 일 상

Asymptotic Analysis of Machine Learning Algorithms

기계 학습 알고리즘의 점근 성질 연구

지도교수 김 용 대

이 논문을 이학박사 학위논문으로 제출함

2020년 2월

서울대학교 대학원

통계학과

온 일 상

온일상의 이학박사 학위논문을 인준함

2020년 2월

위 원 장	**박 병 욱**	**(인)**
부위원장	**김 용 대**	**(인)**
위 원	**장 원 철**	**(인)**
위 원	**원 중 호**	**(인)**
위 원	**채 민 우**	**(인)**

Asymptotic Analysis of Machine Learning Algorithms

By

Ilsang Ohn

A thesis
submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Statistics

Department of Statistics
College of Natural Sciences
Seoul National University

February, 2020

ABSTRACT

Asymptotic Analysis of Machine Learning Algorithms

Ilsang Ohn

Department of Statistics
The Graduate School
Seoul National University

In this thesis, we study the asymptotic properties of three machine learning algorithms including two supervised learning algorithms with deep neural networks and a Bayesian learning method for high-dimensional factor models.

The first research problem involves learning deep neural network (DNN) classifiers. We derive the fast convergence rates of a DNN classifier learned using the hinge loss. We consider various cases for a true probability model and show that the DNN classifier achieves fast convergence rates for all cases, provided its architecture is carefully selected.

The second research topic is learning sparse DNNs. We propose a sparse learning algorithm, which minimizes penalized empirical risk using a novel sparsity-inducing penalty. We establish an oracle inequality for the excess risk of the proposed sparse DNN estimator and derive convergence rates for several learning tasks. In particular, the proposed sparse DNN estimator can adaptively attain minimax optimal convergence rates for nonparametric regression problems.

The third part of the thesis is devoted to Bayesian non-parametric learning for high-dimensional factor models. We propose a prior distribution based on the two-parameter Indian buffet process, which is computationally

tractable. We proved that the resulting posterior distribution concentrates on the true factor dimensionality as well as contracts to the true covariance matrix at a near-optimal rate.

Keywords: Bayesian nonparametrics, Deep neural networks, Fast convergence rates, Factor models, Minimax optimality, Posterior contraction rates, Sparsity

Student Number: 2014-21216

To Sujin

For always being by my side on this journey

Contents

List of Figures

List of Tables

Introduction

0.1 Motivation

The fundamental goal of statistics is to find an optimal machine learning algorithm for a given statistical problem. Unfortunately, there are very few statistical problems for which an indisputable optimal learning algorithms exists. This is partly because the finite sample behavior of a machine learning algorithm is very complex and almost impossible to obtain in general. Thus, comparing machine learning algorithms is usually based on asymptotic theory, which deals with the large-sample behavior of machine learning algorithms. Asymptotic theory provides simplified optimality criteria which are relatively easy to derive. For example, although the risk of a machine learning algorithm is intractable for a finite sample, one can compute the convergence rate of the risk to the minimum risk as the sample size n goes to infinity, and compare that convergence rate with those of other machine learning algorithms.

Classical asymptotic theory focuses on a statistical model in which the sample size n goes to infinity while the number of parameters of the model is fixed as finite. The standard laws of large numbers and the central limit theorem are examples of classical asymptotic theory. These two theoretical statements ensure consistency and asymptotic efficiency, respectively, of classical statistical estimators. However, classical asymptotic theory fails to deal with infinite or high-dimensional statistical models which are very often used to analyze data sets arising in the modern "big data" era. For instance, classical estimators are inconsistent in the high-dimensional regime, where the number of variables is substantially larger than the sample size. This phenomenon requires us to develop new asymptotic theories as well as new machine learning algorithms.

This thesis is devoted to studying asymptotic properties of machine learning algorithms for infinite-dimensional statistical models. The next section describes the research questions and contributions of this thesis.

0.2 Outline and contributions

In this thesis, we study the asymptotic properties of three machine learning algorithms, two of which are supervised learning algorithms with deep neural networks (DNNs), and the other is a Bayesian learning method for high-dimensional factor models.

In Chapter 1, we study the convergence rate of excess 0-1 risk for a deep neural network classifier. In classification problems, it is well known that, when the Tsybakov noise condition is assumed, there exist classifiers attaining fast convergence rates of the excess 0-1 risk, i.e., rates faster than the parametric rate $n^{-1/2}$. We have considered three cases for a true model: (1) the class probability function is smooth; (2) the decision boundary is smooth; and (3) the concentration of the input distribution near the smooth decision boundary is low. We demonstrate that the DNN classifier obtained by minimizing the empirical hinge risk attains fast convergence rates in all three cases. An important implication of this study is that even if we use the hinge loss which is continuous and thus easy to optimize instead of the 0-1 loss, deep learning produces a rate-optimal classifier in each of the two cases. In the third case, we conduct a novel approximation error analysis, which showed that the Bayes classifier is exactly recovered by a DNN function except for the area near the decision boundary. Combined with the assumption of the concentration of the input distribution near the decision boundary, this approximation leads to a convergence rate, which is rather insensitive to the input dimension.

In Chapter 2, we propose a new penalized estimation method for sparse DNNs. A number of empirical observations show that sparse DNNs can dramatically reduce computation time and memory without appreciably harming prediction power. Furthermore, recent theoretical studies proved that DNN estimators obtained by minimizing empirical risk with a certain sparsity constraint can attain optimal convergence rates for regression and classification problems. However, the empirical risk minimizer is nonadaptive and hard to implement due to the discrete nature of its optimization.

In this study, we propose a novel penalized estimation method for sparse DNNs which overcomes these problems. We establish an oracle inequality for the excess risk of the proposed sparse DNN estimator and derive convergence rates for several learning tasks. In particular, the estimator can adaptively attain minimax convergence rates for various nonparametric regression problems. We develop an efficient and scalable gradient-based optimization algorithm.

In Chapter 3, we have studied a consistent Bayesian estimation of the factor dimensionality. A major difficulty in deriving the posterior consistency of the factor dimensionality lies in the presence of "nonsignificant" factors. Without further restrictions on the factor model (e.g., the orthogonal constraint), additional nonsignificant factors, which make a minor change in the spectrum of the covariance matrix, easily appear. Therefore, it is difficult to distinguish the models with different factor dimensionalities by comparing their likelihoods. We propose a novel prior distribution to resolve this issue. The proposed prior is based on the two-parameter Indian buffet process which is computationally tractable. We prove that the resulting posterior distribution concentrates on the true factor dimensionality as well as contracts to the true covariance matrix at the near-optimal rate.

This thesis has two appendices. In Appendix A, we investigate the approximation ability of DNNs with a broad class of activation functions which includes most of the frequently used activation functions. We derive the required depth, width and sparsity of a DNN to approximate any Hölder smooth function up to a given approximation error for the general activation functions. Based on our approximation error analysis, we derive the minimax optimality of the deep neural network estimators with the general activation functions for both regression and classification problems.

In Appendix B, we consider a distribution of a random binary matrix with an infinite number of columns, called a Poisson mixture of finite feature models (PFM). Although the PFM is the most natural prior distribution with which to treat a latent feature model, its posterior computation relies on the reversible jump Markov chain Monte Carlo, which often suffers from slow mixing. We provide different probabilistic representations of the PFM,enabling us to construct a straightforward Gibbs sampling algorithm. As an application, we use the PFM as the prior distribution on the factor loading matrix for Bayesian estimation of a sparse factor model. We prove the posterior consistency of the factor dimensionality and derive the

near-optimal posterior contraction rate of the covariance matrix.

Chapter 1

Fast convergence rates of deep neural networks for classification

1.1 Introduction

Deep learning has received much attention for dimension reduction and classification of objects, such as images, speech, and language. Various supervised and unsupervised deep learning architectures have been developed and applied to large scale real data with great success. Theoretical explanations regarding the success of deep learning have been recently studied. Many researchers have demonstrated that deep neural networks (DNNs) are much more efficient in representing certain complex functions than their shallow counterparts [69, 77, 25], which has been reconfirmed by [100] and [76], who showed that DNNs can approximate a large class of functions, including even discontinuous functions with a parsimonious number of parameters. In turn, using this efficient approximation property of a DNN, Schmidt-Hieber [80], Bauer and Kohler [7] and Imaizumi and Fukumizu [46] proved that, for regression problems, we can estimate a complex function including a discontinuous function using a DNN with the (in the minimax sense) optimal convergence rate. A surprising result is that any linear estimators, which include the kernel ridge estimator, are sub-optimal in estimating a discontinuous function while the DNN is optimal.

In this chapter, we consider classification problems. It is known that there is a classification algorithm that can achieve fast convergence rates of the misclassificaation risk under the Tsybakov low noise condition [63,

93, 94, 3]. Mammen and Tsybakov [63], Tsybakov [93] and Tsybakov and van de Geer [94] considered estimating the classifier by minimizing the empirical misclassification risk, which is computationally infeasible due to the discreteness of the 0-1 loss. Under the smoothness assumption on the conditional class probability, Audibert and Tsybakov [3] estimated the conditional class probability using a local polynomial estimator and obtained a plug-in classifier. Finding the best plug-in classifier, however, requires searching in a given sieve, which is computationally demanding. In contrast, learning a DNN is relatively straightforward owing to the gradient descent algorithm, despite a risk of arriving at bad local minima. We prove that the estimation of a classifier based on the DNN with the hinge loss can achieve fast convergence rates under various situations.

We consider three cases regarding a true classifier: (1) a smooth boundary, (2) smooth conditional class probability, and (3) the margin condition (i.e., the probability of the inputs near the decision boundary is small). We prove that the DNN classifier can achieve fast convergence rates for all of these three cases if the architecture (i.e., the number of layers, number of nodes, and sparsity of the weights) of the DNN is carefully selected. In particular, the DNN classifier is minimax optimal for a smooth decision boundary or a smooth conditional class probability, and achieves faster convergence rates under the margin condition. To the best of the authors' knowledge, no other estimator achieves fast convergence rates for these three cases simultaneously.

The cross-entropy is the standard objective function used in learning a DNN, and is an empirical risk with respect to the logistic loss (i.e., the negative log-likelihood of the logistic model). Learning a DNN with the logistic loss performs quite well in practice. We justify the use of the logistic loss in learning a DNN by showing that the corresponding classifier also achieves a fast convergence rate under certain conditions on the underlying distribution. By small experiments, we illustrate that these assumptions are reasonable for image recognition.

The remainder of this chapter is organized as follows. Section 1.2 describes the hinge loss and DNN classifier. Section 1.3 derives the convergence rates of the excess risk of the DNN classifier for the aforementioned three cases regarding a true model. The fast convergence rate of the DNN classifier with the cross-entropy is derived in Section 1.5, and concluding remarks follow in Section 1.6. All the proofs are gathered in Section 1.7

1.1.1 Notation

We denote by $\mathbb{1}(\cdot)$ the indicator function. Let $\mathbb{R}$ be the set of real numbers and $\mathbb{N}$ be the set of natural numbers. For $m \in \mathbb{N}$, we let $[m] := \{1, \ldots, m\}$. For two given sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ of real numbers, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all sufficiently large n. In addition, we write $a_b \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

For a function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ denotes the domain of the function, let $\|f\|_\infty := \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|$. For a given subset B of $\mathcal{X}$, we let $\|f\|_{\infty,B} := \sup_{\mathbf{x}\in B} |f(\mathbf{x})|$.

Let $\mathbf{m} = (m_1, \ldots, m_d) \in \mathbb{N}_0^d$ be a multiple index, where $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. We define $|\mathbf{m}| := m_1 + \cdots + m_d$ and $\mathbf{x}^{\mathbf{m}} := x_1^{m_1} \cdots x_d^{m_d}$ for a multiple index $\mathbf{m}$. For $f : \mathcal{X} \to \mathbb{R}$ and $\mathbf{m} \in \mathbb{N}_0^d$, let

$$\partial^{\mathbf{m}} f := \frac{\partial^{|\mathbf{m}|} f}{\partial \mathbf{x}^{\mathbf{m}}} := \frac{\partial^{|\mathbf{m}|} f}{\partial x_1^{m_1} \cdots \partial x_d^{m_d}}, \tag{1.1.1}$$

and for $s \in (0, 1]$, let

$$[f]_{s,\mathcal{X}} := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}, \mathbf{x}\neq\mathbf{y}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|^s}.$$

We denote by $\mathcal{C}^m(\mathcal{X})$ and $m \in \mathbb{N}$, the space of m times differentiable functions on $\mathcal{X}$ whose partial derivatives of order $\mathbf{m}$ with $|\mathbf{m}| \leq m$ are continuous. The Hölder space of order α is defined as

$$\mathcal{H}^\alpha(\mathcal{X}) = \left\{ f \in \mathcal{C}^{\lfloor\alpha\rfloor}(\mathcal{X}) : \|f\|_{\mathcal{H}^\alpha(\mathcal{X})} < \infty \right\},$$

where $\|f\|_{\mathcal{H}^\alpha(\mathcal{X})}$ denotes the Hölder norm defined by

$$\|f\|_{\mathcal{H}^\alpha(\mathcal{X})} := \max_{|\mathbf{m}|\leq\lfloor\alpha\rfloor} \|\partial^{\mathbf{m}} f\|_{\infty,\mathcal{X}} + \max_{|\mathbf{m}|=\lfloor\alpha\rfloor} [\partial^{\mathbf{m}} f]_{\alpha-\lfloor\alpha\rfloor,\mathcal{X}}. \tag{1.1.2}$$

We let

$$\mathcal{H}^{\alpha,r}(\mathcal{X}) := \left\{ f \in \mathcal{C}^{\lfloor\alpha\rfloor}(\mathcal{X}) : \|f\|_{\mathcal{H}^\alpha(\mathcal{X})} \leq r \right\},$$

which is a closed ball in the Hölder space of radius r with respect to the Hölder norm.

1.2 Estimation of the classifier with DNNs

We consider a binary classification problem. The (training) data are given as $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, where $\mathbf{X}_i \in \mathcal{X} \subset \mathbb{R}^d$ are input vectors, and $Y_i \in \{-1, 1\}$ are class labels. Here, for simplicity, we set $\mathcal{X} = [0,1]^d$; however, this can be extended to any compact subset of $\mathbb{R}^d$. We assume that $(\mathbf{X}_i, Y_i)$ are independent copies of a random vector $(\mathbf{X}, Y) \sim \mathsf{P}$ for a certain probability measure P. We let P_X be the marginal distribution of $\mathbf{X}$ induced by the joint distribution P and call input distribution. We denote the conditional class probability by $\eta(\cdot)$, that is, we let

$$\eta(\mathbf{x}) := \mathsf{P}(Y = 1 | \mathbf{X} = \mathbf{x}). \tag{1.2.1}$$

1.2.1 About the hinge loss

Before going further, we will first explain technical advantages of the hinge loss to derive fast convergence rates.

For a given real valued function f defined on $\mathcal{X}$, we consider the classifier C_f as $C_f(\mathbf{x}) = \mathrm{sign} f(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$. Then, the *0-1 risk* or simply *risk* of f is defined as

$$\mathcal{E}(f) := \mathsf{E}\left[C_f(\mathbf{x}) \neq Y\right] = \mathsf{E}\left[\mathbb{1}(Y f(\mathbf{x}) < 0)\right],$$

where $\mathbb{1}(\cdot)$ is 1 if $(\cdot)$ is true, and is 0 otherwise. Let C^* be the Bayes classifier which is defined as

$$C^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{E}(f),$$

where $\mathcal{F}$ denotes the set of all measurable real-valued functions on $\mathcal{X}$.

One of the well-studied approaches to estimate C^* is the empirical risk minimization which estimates C^* by $C_{\hat{f}}$, where $\hat{f}$ is the minimizer of the

empirical 0-1 risk defined by

$$\mathcal{E}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i f(\mathbf{X}_i) < 0),$$

over $\mathcal{F}_n$. Here $\mathcal{F}_n$ denotes a given class of real-valued functions (i.e., a sieve) depending on the sample size n.

The empirical risk minimization procedure is theoretical optimal in many cases [e.g., 63, 93, 3], in practice, however, such a procedure is not computationally feasible because minimizing the empirical risk with the 0-1 loss is NP hard [6]. An alternative approach is to replace the 0-1 loss with other computationally easier losses so-called surrogate losses. For a given surrogate loss ϕ, we estimate C^* by $C_{\hat{f}_{\phi,n}}$, where $\hat{f}_{\phi,n}$ is the minimizer of the *empirical ϕ-risk* (or *empirical surrogate risk*) defined as

$$\mathcal{E}_{\phi,n}(f) := \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(\mathbf{X}_i))$$

over $\mathcal{F}_n$. Note that the empirical ϕ-risk converges to the population ϕ-risk defined as

$$\mathcal{E}_\phi(f) := \mathsf{E}(\phi(Y f(\mathbf{X})))$$

and so we expect that the *excess ϕ-risk* (or *excess surrogate risk*)

$$\mathcal{E}_\phi(\hat{f}_{\phi,n}, f^*_\phi) := \mathcal{E}_\phi(\hat{f}_{\phi,n}) - \mathcal{E}_\phi(f^*_\phi)$$

is small, where $f^*_\phi = \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{E}_\phi(f)$. However, the classification performance of $\hat{f}_{\phi,n}$ is measured by the *excess risk* defined as

$$\mathcal{E}(\hat{f}_{\phi,n}, C^*) := \mathcal{E}(\hat{f}_{\phi,n}) - \mathcal{E}(C^*).$$

In general, the fast convergence rate of the excess ϕ-risk does not always imply the fast convergence rate of the excess risk.

Zhang [103] and Bartlett et al. [6] proved that if the surrogate loss ϕ is Fisher consistent (i.e., $\operatorname{sign}(f^*_\phi(\mathbf{x})) = C^*(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$), the following

inequality between the excess risk and excess ϕ-risk holds: there exist constants $C > 0$ and $\rho \in (0, 1]$, which depend on the surrogate loss ϕ, such that

$$\mathcal{E}(f, C^*) \leq C\mathcal{E}_\phi(f, f_\phi^*)^\rho \tag{1.2.2}$$

for any measurable real-valued function f. Moreover, Bartlett et al. [6] showed that $\rho = 1$ for the hinge loss and so the convergence rate of the excess risk can be derived directly from the convergence rate of the excess ϕ-risk. This tool kit could not be applied to the logistic loss since $\rho < 1$.

Another advantage of the hinge loss is that the minimizer of the hinge risk is the Bayes classifier itself, that is, $f_\phi^* = C^*$. This property of the hinge loss makes it possible to analyze the behaviour of the excess risk only with some conditions on the Bayes classifier such as the smooth decision boundary condition assumed by [63, 93, 94]. In contrast, for the logistic loss we have $f_\phi^*(\mathbf{x}) = \log(\eta(\mathbf{x})/(1 - \eta(\mathbf{x})))$ which is the monotone transformation of the conditional class probability function $\eta(\mathbf{x}) := \mathsf{P}(Y = 1|\mathbf{x} = \mathbf{x})$ [35], and hence we need some conditions on the conditional class probability function η which is a larger object than the decision boundary.

1.2.2 Learning DNN with the hinge loss

We consider DNNs that take d-dimensional inputs and produce one-dimensional outputs. A DNN with L many layers, and $\{N^{(l)}, l \in [L]\}$ many nodes at each layer, is defined as

$$z_j^{(l)}(\mathbf{x}) = b_j^{(l)} + \sum_{k=1}^{N^{(l-1)}} W_{j,k}^{(l)} h_k^{(l-1)}(\mathbf{x})$$

and

$$h_j^{(l)}(\mathbf{x}) = \sigma(z_j^{(l)}(\mathbf{x}))$$

for $l = 1, \ldots, L$ and

$$f(\mathbf{x}) = b^{(L+1)} + \sum_{k=1}^{N^{(L)}} W_{1,k}^{(L+1)} h_k^{(L)}(\mathbf{x})$$

with $N^{(0)} = d$ and $h_k^{(0)}(\mathbf{x}) = x_k$. We consider the ReLU activation function $\sigma(z) = (z)_+$. We denote $f(\mathbf{x})$ as $f(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = ((\mathbf{W}^{(l)}, \mathbf{b}^{(l)}))_{l\in[L+1]}$ is the parameter set including all weights and biases.

For the given $\boldsymbol{\theta}$, let $L(\boldsymbol{\theta})$ be the number of layers in $\boldsymbol{\theta}$. Let $N_{\max}(\boldsymbol{\theta})$ be the maximum number of nodes, that is, $f(\cdot|\boldsymbol{\theta})$ has at most $N_{\max}(\boldsymbol{\theta})$ nodes at each layer. We define $\|\boldsymbol{\theta}\|_0$ as the number of nonzero parameters in Θ,

$$\|\boldsymbol{\theta}\|_0 := \sum_{l=1}^{L+1} \left(\|\mathrm{vec}(\mathbf{W}^{(l)})\|_0 + \|\mathbf{b}^{(l)}\|_0 \right)$$

where $\mathrm{vec}(\mathbf{W}^{(l)})$ transforms the matrix $\mathbf{W}^{(l)}$ into the corresponding vector by concatenating the column vectors. Similarly, we define $\|\boldsymbol{\theta}\|_\infty$ as the largest absolute value of the parameters in Θ,

$$\|\boldsymbol{\theta}\|_\infty := \max \left\{ \max_{1\le l\le L+1} \|\mathrm{vec}(\mathbf{W}^{(l)})\|_\infty, \max_{1\le l\le L+1} \|\mathbf{b}^{(l)}\|_\infty \right\}.$$

For a given n, we consider the class of DNNs such that

$$\begin{aligned}&\mathcal{F}^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, F_n)\\ &:= \left\{ f(\cdot|\boldsymbol{\theta}) : L(\boldsymbol{\theta}) \le L_n, N_{\max}(\boldsymbol{\theta}) \le N_n, \|\boldsymbol{\theta}\|_0 \le S_n, \|\boldsymbol{\theta}\|_\infty \le B_n, \|f(\cdot|\boldsymbol{\theta})\|_\infty \le F_n \right\}\end{aligned}$$

where the positive constants L_n, N_n, S_n, B_n, and F_n are specified later.

We let $\hat{f}_{\phi,n}^{\mathrm{DNN}}$ be the minimizer of $\mathcal{E}_{\phi,n}(f)$ over $\mathcal{F}^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, F_n)$ for a given surrogate loss ϕ, i.e.,

$$\hat{f}_{\phi,n}^{\mathrm{DNN}} = \operatorname*{argmin}_{f\in\mathcal{F}^{\mathrm{DNN}}(L_n,N_n,S_n,B_n,F_n)} \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(\mathbf{X}_i)). \tag{1.2.3}$$

In the following section, we prove the fast convergence rates of $\hat{f}_{\phi,n}^{\mathrm{DNN}}$ for various cases of the true model when ϕ is the hinge loss and L_n, N_n, S_n, B_n, and F_n are carefully selected. For detailed formulas of L_n, N_n, S_n, B_n, and F_n in terms of the sample size n, see the proofs of the corresponding theorems in Section 1.7.

1.3 Fast convergence rates of DNN classifiers with the hinge loss

In this section, we consider the hinge loss and derive the convergence rates of the excess risk of $\hat{f}_{\phi,n}^{\mathrm{DNN}}$.

Throughout this chapter, we always assume the Tsybakov noise condition [63, 93].

Assumption N. There exist $c_{\mathrm{N}} > 0$ and $q \in [0, \infty]$ such that for any $t > 0$

$$\mathsf{P}\left(\{\mathbf{X} : |2\eta(\mathbf{X}) - 1| \leq t\}\right) \leq c_{\mathrm{N}} t^q. \tag{1.3.1}$$

We call the constant q in (1.3.1) the *noise exponent*.

We consider three cases regarding a true model: (1) smooth class conditional probability, (2) a smooth decision boundary, and (3) the margin condition. We derive fast convergence rates of the DNN classifier using the hinge loss for all three cases.

1.3.1 Case 1: Smooth conditional class probability

We first assume that $\eta(\mathbf{x})$ is smooth. That is, $\eta(\cdot) \in \mathcal{H}^{\beta,r}([0,1]^d)$ for some $\beta > 0$ and $r > 0$. The following theorem provides the convergence rate of the excess risk of the DNN classifier.

Theorem 1.3.1. *Let $\mathcal{P}_q^{\mathrm{N}}$ be a set of distributions on $[0,1]^d \times \{-1,1\}$ satisfying Assumption N with the noise exponent $q \in [0,\infty]$. If the surrogate loss ϕ is the hinge loss, the classifier $\hat{f}_{\phi,n}^{\mathrm{DNN}}$ defined by (1.2.3) with $F_n = 1$ and carefully selected L_n, N_n, S_n and B_n satisfies*

$$\sup_{\mathsf{P} \in \mathcal{P}_q^{\mathrm{N}} : \eta \in \mathcal{H}^{\beta,r}} \mathsf{E}\left[\mathcal{E}(\hat{f}_{\phi,n}^{\mathrm{DNN}}, C^*)\right] \lesssim \left(\frac{\log^3 n}{n}\right)^{\frac{\beta(q+1)}{\beta(q+2)+d}}, \tag{1.3.2}$$

where the expectation is taken over the training data.

Proof. See Section 1.7.4. □

Audibert and Tsybakov [3] showed that when $\eta(\cdot) \in \mathcal{H}^{\beta}([0,1]^d)$, the minimax lower bound of the excess risk is given by

$$\inf_{\hat{f}_n} \sup_{\mathsf{P} \in \mathcal{P}_q^{\mathsf{N}}: \eta \in \mathcal{H}^{\beta,r}} \mathsf{E}\left[\mathcal{E}(\hat{f}_n, C^*)\right] \gtrsim n^{-\frac{\beta(q+1)}{\beta(q+2)+d}}.$$

Hence, the convergence rate (1.3.2) is minimax optimal up to a logarithmic factor.

1.3.2 Case 2: Smooth boundary

In this case, we impose the smoothness on the decision boundary not on the conditional class probability function. To this, we introduce the notion of piecewise constant functions with smooth boundaries. We adopt the notations and definitions from [76] and [46]. For $g \in \mathcal{H}^{\alpha,r}([0,1]^{d-1})$ and $j \in [d]$, we define a *horizon function* $\Psi_{g,j} : [0,1]^d \to \{0,1\}$ as

$$\Psi_{g,j}(\mathbf{x}) := \mathbb{1}(x_j \geq g(\mathbf{x}_{-j})),$$

where $\mathbf{x}_{-j} := (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)$. For each horizon function, we define the corresponding *basis piece* $I_{g,j}$ as

$$I_{g,j} := \left\{\mathbf{x} \in [0,1]^d : \Psi_{g,j}(\mathbf{x}) = 1\right\}.$$

We define a *piece* by the intersection of K basis pieces. The set of pieces is denoted by

$$\mathcal{A}^{\alpha,r,K} := \left\{ A \subset [0,1]^d : A = \bigcap_{k=1}^{K} I_{g_k,j_k}, g_k \in \mathcal{H}^{\alpha,r}([0,1]^{d-1}), j_k \in [d] \right\}.$$

Let $\mathcal{C}^{\alpha,r,K,T}$ be the set of classifiers of the form

$$C(\mathbf{x}) = 2\sum_{t=1}^{T} \mathbb{1}(\mathbf{x} \in A_t) - 1,$$

for $T \in \mathbb{N}$ and disjoint subsets $A_1, \dots, A_T$ of $\mathcal{X}$ in $\mathcal{A}^{\alpha,r,K}$. In this subsection, we assume that the Bayes classifier belongs to $\mathcal{C}^{\alpha,r,K,T}$.

If most of the data are very close to the decision boundary, any estimator fails to learn the decision boundary rightly. To prevent this situation, we additionally assume that the input distribution has a uniformly bounded density.

Assumption D. The input distribution P_X admits a density p_X with respect to Lebesgue measure and p_X is uniformly bounded.

The following theorem proves the convergence rate of the excess risk of the DNN classifier with the hinge loss.

Theorem 1.3.2. *Let $\mathcal{P}_q^{\mathrm{N,D}}$ be a set of distributions on $[0,1]^d \times \{-1,1\}$ satisfying Assumption N with the noise exponent $q \in [0,\infty]$ and Assumption D. If the surrogate loss ϕ is the hinge loss, the classifier $\hat{f}_{\phi,n}^{\mathrm{DNN}}$ defined by (1.2.3) with $F_n = 1$ and carefully selected L_n, N_n, S_n and B_n satisfies*

$$\sup_{\mathsf{P} \in \bar{\mathcal{P}}_q^{\mathrm{N,D}}: C^* \in \mathcal{C}^{\alpha,r,K,T}} \mathsf{E}\left[\mathcal{E}(\hat{f}_{\phi,n}^{\mathrm{DNN}}, C^*)\right] \lesssim \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha(q+1)}{\alpha(q+2)+(d-1)(q+1)}}, \tag{1.3.3}$$

where the expectation is taken over the training data.

Proof. See Section 1.7.5. □

Tsybakov [93] showed that the minimax lower bound is given by

$$\inf_{\hat{f}_n} \sup_{\mathsf{P} \in \mathcal{P}_q^{\mathrm{N,D}}: C^* \in \mathcal{C}^{\alpha,r,1,1}} \mathsf{E}\left[\mathcal{E}(\hat{f}_n, C^*)\right] \gtrsim n^{-\frac{\alpha(q+1)}{\alpha(q+2)+(d-1)q}}, \tag{1.3.4}$$

where the infimum is taken over all classifiers $\hat{f}_n : ([0,1]^d \times \{-1,1\})^n \mapsto \mathcal{F}$ and $\mathcal{F}$ is a set of all measurable functions. Unfortunately, the convergence rate of (1.3.3) is not minimax optimal. However, we can improve the convergence rate further by assuming an additional condition on the conditional class probability $\eta(\mathbf{x})$ as is done by [94].

Assumption R. For the Bayes classifier which is given by

$$C^*(\mathbf{x}) = 2\sum_{t=1}^{T} \mathbb{1}\left(\mathbf{x} \in \bigcap_{k=1}^{K} I_{g_{t,k}^*, j_{t,k}^*}\right) - 1,$$

there exist $\epsilon_0 > 0$ and $c_R > 0$ such that for any $\epsilon \in (0, \epsilon_0]$,

$$\sup_{C \in \mathcal{C}(\epsilon, C^*)} \mathcal{E}(C, C^*) \leq c_R \epsilon^{(q+1)/q}, \tag{1.3.5}$$

where $q \in [0, \infty]$ is the noise exponent and $\mathcal{C}(\epsilon, C^*)$ is the set of classifiers defined by

$$\mathcal{C}(\epsilon, C^*) := \left\{ 2\sum_{t=1}^{T} \mathbb{1}\left(\mathbf{x} \in \cap_{k=1}^{K} I_{g_{t,k}, j_{t,k}^*}\right) - 1 : \max_{t \in [T]} \max_{k \in [K]} \|g_{t,k} - g_{t,k}^*\|_\infty \leq \epsilon \right\}.$$

Remark 1.3.1. By the property of the symmetric difference operator $\triangle$, we have that for $C(\mathbf{x}) = 2\sum_{t=1}^{T} \mathbb{1}\left(\mathbf{x} \in \cap_{k=1}^{K} I_{g_{t,k}, j_{t,k}^*}\right) - 1$,

$$\begin{aligned} \mathcal{E}(C, C^*) &\leq \mathsf{P}_X\left(\cap_{k=1}^{K} I_{g_{t,k}, j_{t,k}^*} \triangle \cap_{k=1}^{K} I_{g_{t,k}^*, j_{t,k}^*}\right) \\ &\leq \sum_{t=1}^{T}\sum_{k=1}^{K} \mathsf{P}_X\left(I_{g_{t,k}, j_{t,k}^*} \triangle I_{g_{t,k}^*, j_{t,k}^*}\right). \end{aligned}$$

By Assumption D, which allows us to interchange P_X and Lebesgue measure, there is a constant $c_1 > 0$ such that for every $t \in [T]$ and $k \in [K]$,

$$\mathsf{P}_X\left(I_{g_{t,k}, j_{t,k}^*} \triangle I_{g_{t,k}^*, j_{t,k}^*}\right) \leq c_1 \|g_{t,k} - g_{t,k}^*\|_1.$$

That is, Assumption D provides the looser upper bound of the excess risk given below than the one given in (1.3.5):

$$\sup_{C \in \mathcal{C}(\epsilon, C^*)} \mathcal{E}(C, C^*) \leq c_R \epsilon.$$

Remark 1.3.2. Tarigan and Van De Geer [88] provided the following sufficient condition for Assumption R when $T = 1$ and $K = 1$: there is a constant

$c_0 > 1$ such that

$$|2\eta(\mathbf{x}) - 1| \leq c_0 |x_j - g^*(\mathbf{x}_{-j})|^{1/q} \tag{1.3.6}$$

for any $\mathbf{x} \in [0,1]^d$. This condition imposes a certain smoothness on $\eta(\mathbf{x})$ near the decision boundary. We generalizes (1.3.6) for $T \geq 1$ and $K \geq 1$ as

$$|2\eta(\mathbf{x}) - 1| \leq c_0 \min_{t\in[T]} \min_{k\in[K]} \left| x_{j^*_{t,k}} - g^*_{t,k}(\mathbf{x}_{-j^*_{t,k}}) \right|^{1/q} \tag{1.3.7}$$

for any $\mathbf{x} \in [0,1]^d$. It can be shown that Assumption R holds under (1.3.7) similarly to the proof of Lemma 4.1. of [88].

The following theorem proves the minimax convergence rate of the DNN classifier under Assumption R.

Theorem 1.3.3. *Let $\mathcal{P}_q^{\mathrm{N,R}}$ be a set of distributions on $[0,1]^d \times \{-1,1\}$ satisfying Assumption N with the noise exponent $q \in [0,\infty]$ and Assumption R. If the surrogate loss ϕ is the hinge loss, the classifier $\hat{f}^{\mathrm{DNN}}_{\phi,n}$ defined by (1.2.3) with $F_n = 1$ and carefully selected L_n, N_n, S_n and B_n satisfies*

$$\sup_{\mathsf{P}\in\mathcal{P}_q^{\mathrm{N,R}}:C^*\in\mathcal{C}^{\alpha,r,K,T}} \mathsf{E}\left[\mathcal{E}(\hat{f}^{\mathrm{DNN}}_{\phi,n}, C^*)\right] \lesssim \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha(q+1)}{\alpha(q+2)+(d-1)q}}, \tag{1.3.8}$$

where the expectation is taken over the training data.

Proof. See Section 1.7.6. □

Now, the convergence rate (1.3.8) is minimax optimal up to a logarithmic factor. The estimators of [93] and [94] also achieved the minimax lower bound, but they considered the empirical 0-1 risk minimizer, which is not computationally feasible.

1.3.3 Case 3: Margin condition

The convergence rate (1.3.8) can be improved if we assume that the density of an input variable is small around the decision boundary. Let

$$D^* := \{\mathbf{x} : \eta(\mathbf{x}) = 1/2\}$$

and

$$\mathrm{dist}(\mathbf{x}, D^*) := \inf_{\mathbf{x}^* \in D^*} \|\mathbf{x} - \mathbf{x}^*\|_2$$

where $\|\cdot\|_2$ denotes the Euclidian norm. We introduce the following condition on the probability measure P_X.

Assumption M. There exist $c_{\mathrm{M}} > 0$, $\epsilon_0 > 0$, and $\gamma \in [1, \infty]$ such that for any $\epsilon \in (0, \epsilon_0]$,

$$\mathsf{P}\left(\{\mathbf{X} : \mathrm{dist}(\mathbf{X}, D^*) \leq \epsilon\}\right) \leq c_{\mathrm{M}} \epsilon^{\gamma}. \tag{1.3.9}$$

We call the constant γ the *margin exponent*.

Assumption M is considered by [84] who proves that the support vector machine with the Gaussian kernel achieves a fast convergence rate under Assumption M. The following theorem proves that a similar convergence rate can be achieved using the DNN classifier.

Theorem 1.3.4. *Let $\mathcal{P}_{q,\gamma}^{\mathrm{N,M}}$ be a set of distributions on $[0,1]^d \times \{-1, 1\}$ satisfying Assumption N with the noise exponent $q \in [0, \infty]$ and Assumption M with the margin exponent $\gamma \in [1, \infty]$. If the surrogate loss ϕ is the hinge loss, the classifier $\hat{f}_{\phi,n}^{\mathrm{DNN}}$ defined by (1.2.3) with $F_n = 1$ and carefully selected L_n, N_n, S_n and B_n satisfies*

$$\sup_{\mathsf{P} \in \mathcal{P}_{q,\gamma}^{\mathrm{N,M}} : C^* \in \mathcal{C}^{\alpha,r,K,T}} \mathsf{E}\left[\mathcal{E}(\hat{f}_{\phi,n}^{\mathrm{DNN}}, C^*)\right] \lesssim \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha(q+1)}{\alpha(q+2)+(d-1)(q+1)/\gamma}}, \tag{1.3.10}$$

where the expectation is taken over the training data.

Proof. See Section 1.7.7. □

Remark 1.3.3. Theorem 1.3.4 does not assume Assumption D. Assuming

Assumption M alone is sufficient to prevent situations where most of the data are very close to the decision boundary.

An interesting feature of the convergence rate (1.3.10) is that the dependency of the input dimension d diminishes as γ increases. In the extreme case where $\gamma \to \infty$, the convergence rate becomes $n^{-(q+1)/(q+2)}$ up to the logarithm factor, which depends on neither the smoothness of the boundary nor the dimension of the input. This partly explains why the DNN classifier works well with high-dimensional inputs such as images.

To investigate the validity of the margin condition, we explore the area near the decision boundary obtained by "5" and "7" characters of the MNIST dataset. We first estimate the density of the dataset by use of the PixelCNN [96]. Then we sample 200 images from the two classes "5" and "7" and generate artificial samples near the decision boundary based on the adversarial training method proposed by [48]. Finally, we compare the log-density values of the sampled real images and artificial images. Figure 1.1 draw some representative real and artificial images and the boxplots of the log-density values. It is obvious that the log-density values of the artificial images are much lower than those of the real images, which suggests that the assumption of a large margin exponent is not too absurd.

Another interesting observation is that some artificial images look real images even though most images are blurred versions of real images. This fact indicates that there are images which do not exist in reality but similar to real images. That is, classification and generation would be quite different subjects.

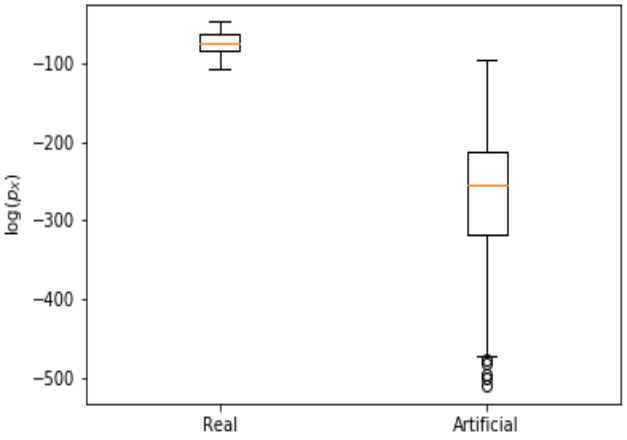
1.4 Adaptive estimation

In practice, we do not know the smoothnees parameter α (or β) of the true decision boundary (or the true conditional class probability) that affects the choice of the DNN architecture parameters L_n, N_n, S_n and B_n. We may select them data-adaptively. For example, a model selection approach can be applied.

For simplicity, we focus on the smooth boundary case with $q = \infty$. For given $\omega > 0$, let $\xi_{n,\omega} = (\log n^3/n)^{\omega/(\omega+d-1)}$, which is the convergence rate of the excess risk of the DNN classifier with the hinge loss when

(A) Real images

(B) Artificial images



(C) Boxplots of the log densities

FIGURE 1.1: Input density estimation for real and artificial images generated near the decision boundary. (a) and (b) are representative samples of real and artificial images, respectively and (c) is the boxplots of the log-densities of the real and artificial images

$C^* \in \mathcal{C}^{\omega,r,K,T}$ given in Theorem 1.3.2. Let $L_{n,\omega}, N_{n,\omega}, S_{n,\omega}$ and $B_{n,\omega}$ be the DNN architecture parameters corresponding to the convergence rate $\xi_{n,\omega}$, which are given by

$$\begin{aligned} L_{n,\omega} &\asymp \log n \\ N_{n,\omega} &\asymp \left(n/\log^3 n\right)^{d-1/(\omega+d-1)} \\ S_{n,\omega} &\asymp \left(n/\log^3 n\right)^{d-1/(\omega+d-1)} \log n \\ B_{n,\omega} &\asymp \left(n/\log^3 n\right)^{\omega/(\omega+d-1)}. \end{aligned}$$

We define $\hat{f}_{n,\omega}$ as

$$\hat{f}_{n,\omega} = \underset{f \in \mathcal{F}_{n,\omega}^{\mathrm{DNN}}}{\operatorname{argmin}} \mathcal{E}_{\phi,n}(f),$$

where

$$\mathcal{F}_{n,\omega}^{\mathrm{DNN}} := \mathcal{F}^{\mathrm{DNN}}(L_{n,\omega}, N_{n,\omega}, S_{n,\omega}, B_{n,\omega}, 1)$$

Finally, let $\mathcal{A}_n$ be the set of candidate smoothness parameters ωs. Then, we estimate $\hat{\omega}$ as

$$\hat{\omega} = \underset{\omega \in \mathcal{A}_n}{\operatorname{argmin}} \left[\mathcal{E}_{\phi,n}(\hat{f}_{n,\omega}) + \mathrm{pen}_n(\omega) \right], \tag{1.4.1}$$

where $\mathrm{pen}_n(\omega)$ is a penalty function. The next theorem states that $\hat{f}_{n,\hat{\omega}}$ with a suitable choice of $\mathcal{A}_n$ and the penalty function can attain the fast convergence rate.

Theorem 1.4.1. *Let $\mathcal{P}_\infty^{\mathrm{N,D}}$ be a set of distributions on $[0,1]^d \times \{-1,1\}$ satisfying Assumption N with the noise exponent $q = \infty$ and Assumption D. Let ϕ be the hinge loss. For a given $\tau > 0$, we set*

$$\mathcal{A}_n := \left\{ \frac{d-1}{\log n/k - 1} : k = 1, \dots, \lfloor \log n \rfloor \right\} \bigcap \left\{ \omega : \frac{1}{\tau} < \omega < \tau \right\}.$$

Let $(z_\omega)_{\omega \in \mathcal{A}_n}$ be a sequence of positive real numbers such that $\sum_{\omega \in \mathcal{A}_n} \mathrm{e}^{-z_\omega} \leq 1$. If we let the penalty function be

$$\begin{aligned} \mathrm{pen}_n(\omega) =& (S_{n,\omega}+1) \frac{20(250)^2 L_{n,\omega} \log\left\{ n(N_{n,\omega}+1)(B_{n,\omega} \vee 1) \right\}}{n} \\ &+ \frac{178\left(z_\omega + \log 2 + 3\log(178)\right)}{3n}, \end{aligned} \tag{1.4.2}$$

then we have

$$\sup_{\mathsf{P} \in \mathcal{P}_\infty^{\mathrm{N,D}}: C^* \in C^{\alpha,r,K,T}} \mathsf{E}\left[\mathcal{E}\left(\hat{f}_{n,\hat{\omega}}, C^* \right) \right] \lesssim \left(\frac{\log^3 n}{n} \right)^{\frac{\alpha}{\alpha+d-1}} \tag{1.4.3}$$

for any $\alpha \in (1/\tau, \tau)$, where the expectation is taken over the training data.

Proof. See Section 1.7.8. □

Remark 1.4.1. We introduce τ in Theorem 1.4.1 to make the proof simpler. We can extend it to $\tau = \infty$ but it requires messy calculations of the constant terms L_0, N_0, S_0, B_0 in Proposition 1.7.7 because those constant terms

depend on ω. By letting $\omega \in (1/\tau, \tau)$, we can let the constant terms uniformly bounded and hence ignore them.

1.5 Use of the logistic loss

In this section, we prove that the convergence rate of the excess risk of the DNN estimator with the logistic loss can be fast when the noise exponent q and margin exponent γ are large. To be more specific, we assume the following two conditions instead of Assumption N and Assumption M:

Assumption N′. There exists a constant $\eta_0 \in (0,1)$ such that

$$\mathsf{P}(\{\mathbf{X} : \left|2\eta(\mathbf{X}) - 1\right| \leq \eta_0\}) = 0.$$

Assumption M′. There exists a constant $m_0 > 0$ such that

$$\mathsf{P}\left(\{\mathbf{X} : \mathrm{dist}(\mathbf{X}, D^*) \leq m_0\}\right) = 0$$

These two conditions are the essentially the same as that $q = \infty$ and $\gamma = \infty$ in Assumption N and Assumption M.

These two conditions are expected to hold in many image recognition problems. The validity of Assumption M′ has been already explained in Section 1.3.3. For Assumption N′, in Figure 1.2 we draw the histogram of the conditional class probabilities of the test data of CIFAR10 data estimated by a convolutional neural network (CNN) with the logistic loss. Most of conditional class probabilities are close to either 0 or 1 and very few are around 0.5, which illustrates that the Assumption N′ is not too strange.

In the following theorem, we derive a fast convergence rate for the DNN classifier with the logistic loss.

Theorem 1.5.1. *Let $\mathcal{P}_{\infty,\infty}^{\mathrm{N,M}}$ be a set of distributions on $[0,1]^d \times \{-1,1\}$ satisfying Assumption N′ and Assumption M′. Let $\phi(z) = \log(1+\exp(-z))$. Then there exist positive constants L_0, N_0, S_0, B_0 and F_0 such that the estimator $\hat{f}_{\phi,n}^{\mathrm{DNN}}$ given by*

$$\hat{f}_{\phi,n}^{\mathrm{DNN}} = \operatorname*{argmin}_{f \in \mathcal{F}_0^{\mathrm{DNN}}} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i f(\mathbf{x}_i)),$$

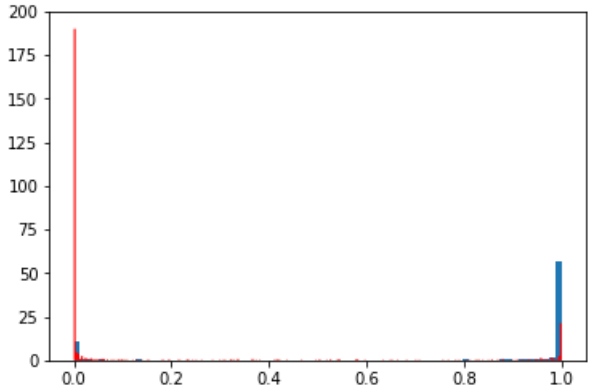


FIGURE 1.2: Histogram of the conditional class probabilities estimated using a DNN with the logistic loss for CIFAR10 data. The blue bins are for the 'dog' samples, and the red bins indicate the 'cat' samples.

with $\mathcal{F}_0^{\text{DNN}} := \mathcal{F}^{\text{DNN}}(L_0, N_0, S_0, B_0, F_0)$ *satisfies*

$$\sup_{\mathsf{P} \in \mathcal{P}_{\infty,\infty}^{\mathrm{N},\mathrm{M}}: C^* \in \mathcal{C}^{\alpha,r,K,T}} \mathsf{E}\left[\mathcal{E}(\hat{f}_{\phi,n}^{\text{DNN}}, C^*)\right] \lesssim \frac{\log^{1+\kappa} n}{n} \tag{1.5.1}$$

for any $\kappa > 0$, *where the expectation is taken over the training data,*

Proof. See Section 1.7.9. □

We compare the performance of the two classifiers learned using the two surrogate losses - the logistic loss and the hinge loss. We analyze three benchmark datasets for image recognition, that is, MNIST, SVHN, and CIFAR10, where for each dataset we select two classes that are most difficult to recognize. The data descriptions and selected classes are summarized in Table 1.1.

For the MNIST dataset, we used a DNN with five hidden layers, whose numbers of nodes were 1200, 600, 300, 150, and 150, respectively. All hidden layers are followed by batch normalization [47]. In addition, for the SVHN and CIFAR10 datasets, we used the CNN models whose architectures are provided in Table 1.2. The Adam is used for optimization with the learning rate 10^{-3}.

TABLE 1.1: Data summary

Data	# of training data	# of test data	Input dim.	Selected classes
MNIST	60,000	10,000	28×28	'5' vs. '7'
SVHN	73,257	26,032	$3 \times 32 \times 32$	'4' vs. '9'
CIFAR10	60,000	50,000	$3 \times 32 \times 32$	'cat' vs. 'dog'

TABLE 1.2: CNN models used in our experiments over SVHN and CIFAR-10. All convolutional (conv.) and fully connected (FC) layers are followed by the batch normalization.

SVHN	CIFAR10
32×32 RGB images	
3×3 conv. 64 ReLU	3×3 conv. 96 ReLU
3×3 conv. 64 ReLU	3×3 conv. 96 ReLU
3×3 conv. 64 ReLU	3×3 conv. 96 ReLU
2×2 max-pool, stride 2	
dropout, $p = 0.5$	
3×3 conv. 128 ReLU	3×3 conv. 192 ReLU
3×3 conv. 128 ReLU	3×3 conv. 192 ReLU
3×3 conv. 128 ReLU	3×3 conv. 192 ReLU
2×2 max-pool, stride 2	
dropout, $p = 0.5$	
3×3 conv. 128 ReLU	3×3 conv. 192 ReLU
1×1 conv. 128 ReLU	1×1 conv. 192 ReLU
1×1 conv. 128 ReLU	1×1 conv. 192 ReLU
global average pool, $6 \times 6 \to 1 \times 1$	
FC $128 \to 1$	FC $192 \to 1$

TABLE 1.3: Test errors of the classifiers learned using the hinge and logistic losses with various training data sizes.

Data	# of training samples per each class	Hinge loss Mean	Hinge loss SE	Logistic loss Mean	Logistic loss SE
MNIST	50	0.9318	0.0078	0.9359	0.0100
	500	0.9806	0.0031	0.9799	0.0024
	5000	0.9929	0.0006	0.9925	0.0005
SVHN	50	0.7877	0.0698	0.7851	0.0798
	500	0.9500	0.0061	0.9545	0.0063
	5000	0.9796	0.0011	0.9801	0.0014
CIFAR10	50	0.6628	0.0123	0.6698	0.0096
	500	0.7758	0.0090	0.7804	0.0081
	5000	0.8760	0.0064	0.8788	0.0047

Table 1.3 summarizes the test data error rates for various sizes of training data. The results are the averages (and standard errors) of 100 randomly selected training data, which amply show that the two estimators compete well with each other.

1.6 Concluding remarks

We showed that a DNN is very flexible in the sense that it achieves fast convergence rates for various cases regarding a true model. It is interesting to note that a DNN is not only good at the case of a smooth decision boundary but also the case of a smooth conditional class probability. In addition, a DNN can fully utilize the margin condition.

We showed that using the logistic loss is promising under the two rather strong conditions. This limitation is mainly due to technical difficulties, and we believe that the logistic loss works well for other cases. We will pursue this issue in near future.

We did not consider a computational issue in this chapter. Learning a DNN with a sparsity constraint has not been fully studied, although various methods have been proposed (e.g., [61], [43], [97], [34] and [62]). A learning algorithm that supports our theoretical results will be worth pursuing.

1.7 Proofs

1.7.1 Complexity of a class of DNNs

In this section, we introduce the complexity measures of a given class of functions which are needed for the proofs. Let $\|\cdot\|_p$ for $1 \leq p < \infty$ be defined as $\|f\|_p := \left(\int_{\mathcal{X}} |f(\mathbf{x})|^p \mathrm{d}\mu(\mathbf{x})\right)^{1/p}$, where μ denotes Lebesgue measure and $\|f\|_\infty := \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|$.

Let $\mathcal{F}$ be a given class of real-valued functions defined on $\mathcal{X}$. Let $\delta > 0$ and $p \in [1, \infty]$. A collection $\{f_i \in \mathcal{F} : i \in [N]\}$ is called a *δ-covering set* of $\mathcal{F}$ with respect to the L_p norm if, for all $f \in \mathcal{F}$, there exists f_i in the collection such that $\|f - f_i\|_p \leq \delta$. The cardinality of the minimal δ-covering set is called the *δ-covering number* of $\mathcal{F}$ with respect to the L_p norm, and is denoted by $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_p)$, that is,

$$\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_p) := \inf\left\{ N \in \mathbb{N} : \exists f_1, \ldots, f_N \text{ such that } \mathcal{F} \subset \bigcup_{i=1}^{N} B_p(f_i, \delta) \right\},$$

where $B_p(f_i, \delta) := \{f \in \mathcal{F} : \|f - f_i\|_p \leq \delta\}$.

A collection of pairs $\{(f_i^L, f_i^U) \in \mathcal{F} \times \mathcal{F} : i \in [N]\}$ is called a *δ-bracketing set* of $\mathcal{F}$ with respect to the L_p norm if $\|f_i^U - f_i^L\| \leq \delta$ for all $i \in [N]$, and for any $f \in \mathcal{F}$, there is a pair (f_i^L, f_i^U) in the collection such that $f_i^L \leq f \leq f_i^U$. The cardinality of the minimal δ-bracketing set is called the *δ-bracketing number* of $\mathcal{F}$ with respect to the L_p norm, and is denoted by $\mathcal{N}_B(\delta, \mathcal{F}, \|\cdot\|_p)$. The *$\delta$-bracketing entropy* denoted by $H_B(\delta, \mathcal{F}, \|\cdot\|_p)$ is the logarithm of the δ-bracketing number, i.e., $H_B(\delta, \mathcal{F}, \|\cdot\|_p) := \log \mathcal{N}_B(\delta, \mathcal{F}, \|\cdot\|_p)$.

For any $\delta > 0$, it is known (see, for example, Lemma 2.1 of [95]) that

$$\log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_p) \leq H_B(\delta, \mathcal{F}, \|\cdot\|_p)$$

for any $p \in [1, \infty)$, and

$$H_B(\delta, \mathcal{F}, \|\cdot\|_p) \leq \log \mathcal{N}(\delta/2, \mathcal{F}, \|\cdot\|_\infty) \tag{1.7.1}$$

provided that $\mu(\mathcal{X}) = 1$.

The following proposition states the upper bound of the δ-entropy of a neural network function space.

Proposition 1.7.1 (Lemma 3 of [87], Lemma 5 of [80]). *For any $\delta > 0$,*

$$\begin{aligned} &\log \mathcal{N}\left(\delta, \mathcal{F}^{\mathrm{DNN}}(L, N, S, B, \infty), \|\cdot\|_\infty\right) \\ &\quad \le 2L(S+1)\log\left(\delta^{-1}(L+1)(N+1)(B \vee 1)\right). \end{aligned} \tag{1.7.2}$$

where $B \vee 1 = \max\{B, 1\}$.

1.7.2 Convergence rate of the excess surrogate risk for general surrogate losses

In this subsection, we derive the convergence rate of the excess ϕ-risk under regularity conditions, which is used repeatedly in the following subsections. The regularity conditions and techniques of the proof are minor modifications of those in [74]; however, we present the complete conditions and proof for the sake of readers' convenience.

We assume the following conditions.

(A1) ϕ is Lipschitz, i.e., there exists a constant $c_1 > 0$ such that $|\phi(z_1) - \phi(z_2)| \le c_1|z_1 - z_2|$ for any $z_1, z_2 \in \mathbb{R}$.

(A2) For a positive sequence $\{a_n\}_{n\in\mathbb{N}}$, there exists a sequence of function classes $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ such that

$$\mathcal{E}_\phi(f_n, f_\phi^*) \le a_n$$

for some $f_n \in \mathcal{F}_n$.

(A3) There exists a sequence $\{F_n\}_{n\in\mathbb{N}}$ with $F_n \gtrsim 1$ such that $\sup_{f\in\mathcal{F}_n} \|f\|_\infty \le F_n$.

(A4) There exists a constant $\nu \in (0,1]$ such that for any $f \in \mathcal{F}_n$ and any $n \in \mathbb{N}$,

$$\mathsf{E}\left[\left\{\phi(Yf(\mathbf{X})) - \phi(Yf_\phi^*(\mathbf{X}))\right\}^2\right] \leq c_2 F_n^{2-\nu}\{\mathcal{E}_\phi(f, f_\phi^*)\}^\nu$$

for a constant $c_2 > 0$ depending only on ϕ and $\eta(\cdot)$.

(A5) There exists a sequence $\{\delta_n\}_{n\in\mathbb{N}}$ such that

$$H_B(\delta_n, \mathcal{F}_n, \|\cdot\|_2) \leq c_3 n \left(\frac{\delta_n}{F_n}\right)^{2-\nu},$$

for some constant $c_3 > 0$, with $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ in (A2), $\{F_n\}_{n\in\mathbb{N}}$ in (A3), and ν in (A4).

For a proof of the general convergence result, we apply the large deviation inequality of [81] presented in Lemma 1.7.2.

Lemma 1.7.2 (Theorem 3 of [81]). *Let $\mathcal{F}$ be the class of functions bounded above by F. Assume that $\mathsf{E}f(Z) = 0$ for any $f \in \mathcal{F}$ and $v \geq \sup_{f\in\mathcal{F}} \mathrm{Var}(f(Z))$ for some $v > 0$. Suppose that there exists $\zeta > 0$ such that*

(C1) $H_B(v^{1/2}, \mathcal{F}, \|\cdot\|_2) \leq \zeta n M^2/(8(4v + MF/3))$,

(C2) $M \leq \zeta v/(4F)$, $v^{1/2} \leq F$,

(C3) if $\zeta M/8 < v^{1/2}$,

$$M^{-1}\int_{\zeta M/32}^{v^{1/2}} H_B(u, \mathcal{F}, \|\cdot\|_2)^{1/2}\mathrm{d}u \leq \frac{n^{1/2}\zeta^{3/2}}{2^{10}}.$$

Then

$$\mathsf{P}^*\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(Z_i) - \mathsf{E}f(Z_i)) \geq M\right) \leq 3\exp\left\{-(1-\zeta)\frac{nM^2}{2(4v + MF/3)}\right\},$$

where P^* *denotes the outer probability measure.*

The following theorem is the main result of this section, which gives the convergence rate of the excess ϕ-risk.

Theorem 1.7.3. *Suppose that the conditions (A1)-(A5) are met. Let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence in (A2) and $\{\delta_n\}_{n\in\mathbb{N}}$ be a sequence in (A5) with $c_3 = 2^{-11-6\nu}c_1^{\nu-2}/\max\{c_2(1+4^\nu), 64(2c_1)^{2-\nu}\}$, where c_1 and c_2 are constants appearing in (A1) and (A4), respectively. Let $\epsilon_n^2 = \max\{2a_n, 2^7c_1\delta_n\}$. Then the empirical ϕ-risk minimizer $\hat{f}_{\phi,n}$ over $\mathcal{F}_n$ satisfies*

$$\mathsf{P}\left(\mathcal{E}_\phi(\hat{f}_{\phi,n}, f_\phi^*) \geq \epsilon_n^2\right) \lesssim \exp(-cn(\epsilon_n^2/F_n)^{2-\nu}),$$

for some universal constant $c > 0$.

Proof. We define the following empirical process

$$Z_n(f) := \frac{1}{n}\sum_{i=1}^n \left[\phi(Y_i f_n(\mathbf{X}_i)) - \phi(Y_i f(\mathbf{X}_i)) - \mathsf{E}\left\{\phi(Y f_n(\mathbf{X})) - \phi(Y f(\mathbf{X}))\right\}\right], \tag{1.7.3}$$

where $f_n \in \mathcal{F}_n$ is a function such that $\mathcal{E}_\phi(f_n, f_\phi^*) \leq a_n$.

Since $\hat{f}_{\phi,n}$ minimizes $\mathcal{E}_{\phi,n}(f) = \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(\mathbf{X}_i))$,

$$\begin{aligned}
&\mathsf{P}\left\{\mathcal{E}_\phi(\hat{f}_{\phi,n}, f_\phi^*) \geq \epsilon_n^2\right\} \\
&\quad\leq \mathsf{P}^*\left(\sup_{f\in\mathcal{F}_n:\mathcal{E}_\phi(f,f_\phi^*)\geq\epsilon_n^2} \frac{1}{n}\sum_{i=1}^n \{\phi(Y_i f_n(\mathbf{X}_i)) - \phi(Y_i f(\mathbf{X}_i))\} \geq 0\right).
\end{aligned}$$

Let us define

$$\mathcal{F}_{n,i} := \{f \in \mathcal{F}_n : 2^{i-1}\epsilon_n^2 \leq \mathcal{E}_\phi(f, f_\phi^*) < 2^i\epsilon_n^2\}.$$

Note that for $i \in \mathbb{N}$ such that $2^{i-1}\epsilon_n^2 > 2c_1F_n$, $\mathcal{F}_{n,i}$ is an empty set. This is because for any $f \in \mathcal{F}_n$, $\|f\|_\infty \leq F_n$, and thus $\mathcal{E}_\phi(f, f_\phi^*) \leq \mathsf{E}|\phi(Yf(\mathbf{X})) - \phi(Yf_\phi^*(\mathbf{X}))| \leq c_1\mathsf{E}|f(\mathbf{X}) - f_\phi^*(\mathbf{X})| \leq 2c_1F_n$. Therefore, $\{f \in \mathcal{F}_n : \mathcal{E}_\phi(f, f_\phi^*) \geq \epsilon_n^2\} \subset \bigcup_{i=1}^{i_n^*} \mathcal{F}_{n,i}$, where $i_n^* = \inf\{i \in \mathbb{N} : 2^{i-1}\epsilon_n^2 > 2c_1F_n\}$. Thus, we only

deal with $\mathcal{F}_{n,i}$ for $i \le i_n^*$. Because $\mathcal{E}_\phi(f_n, f_\phi^*) \le a_n \le \epsilon_n^2/2$, we have

$$\inf_{f\in\mathcal{F}_{n,i}} \mathsf{E}\{\phi(Yf(\mathbf{X})) - \phi(Yf_n(\mathbf{X}))\} = \inf_{f\in\mathcal{F}_{n,i}} \{\mathcal{E}_\phi(f, f_\phi^*) - \mathcal{E}_\phi(f_n, f_\phi^*)\} \ge 2^{i-2}\epsilon_n^2.$$

We introduce the notation $M_{n,i} = 2^{i-2}\epsilon_n^2$ for a concise expression. By the triangle inequality and (A4), we obtain the following variance bound

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_{n,i}} \mathsf{E}\{\phi(Yf(\mathbf{X})) - \phi(Yf_n(\mathbf{X}))\}^2 \\
&\qquad \le c_2 F_n^{2-\nu} \left(\sup_{f\in\mathcal{F}_{n,i}} \mathcal{E}_\phi(f, f_\phi^*)^\nu + \mathcal{E}_\phi(f_n, f_\phi^*)^\nu \right) \\
&\qquad \le c_2(1+4^\nu) F_n^{2-\nu} (2^{i-2}\epsilon_n^2)^\nu \\
&\qquad = c_2(1+4^\nu) F_n^{2-\nu} M_{n,i}^\nu.
\end{aligned} \tag{1.7.4}$$

Now, we have

$$\mathsf{P}\left\{\mathcal{E}_\phi(\hat{f}_n, f_\phi^*) \ge \epsilon_n^2\right\} \le \sum_{i=1}^{i_n^*} \mathsf{P}^* \left(\sup_{f\in\mathcal{F}_{n,i}} Z_n(f) \ge M_{n,i} \right). \tag{1.7.5}$$

To bound the right-hand side of (1.7.5), we apply Lemma 1.7.2 to the class of functions

$$\mathcal{H}_{n,i} := \{(\mathbf{x}, y) \mapsto \phi(yf_n(\mathbf{x})) - \phi(yf(\mathbf{x})) : f \in \mathcal{F}_{n,i}\},$$

with $\zeta = 1/2$, $F = D_1 F_n$, $M = M_{n,i}$, and $v = v_{n,i} = D_2 F_n^{2-\nu} M_{n,i}^\nu$ where we let

$$D_1 := \frac{1}{8(2c_1)^{1-\nu}} D_2, \quad D_2 := \max\{c_2(1+4^\nu), 64(2c_1)^{2-\nu}\}.$$

Note that for any $h \in \mathcal{H}_{n,i}$, $\|h\|_\infty \le c_1\|f_n - f\|_\infty \le 2c_1 F_n$ and $\sup_{h\in\mathcal{H}_{n,i}} \mathrm{Var}(h(\mathbf{X}, Y)) \le c_2(1+4^\nu) F_n^{2-\nu} M_{n,i}^\nu$ by (1.7.4). Since $D_1 \ge 2c_1$ and $D_2 \ge c_2(1+4^\nu)$, we have $\sup_{h\in\mathcal{H}_{n,i}} \|h\|_\infty \le D_1 F_n$ and $\sup_{h\in\mathcal{H}_{n,i}} \mathrm{Var}(h(\mathbf{X}, Y)) \le v_{n,i}$. Now we will check (C1), (C2) and (C3) of Lemma 1.7.2. Because $M_{n,i} \le 2c_1 F_n$ for any

$i \le i_n^*$ and $D_2 \ge 64(2c_1)^{2-\nu}$,

$$\begin{aligned}\frac{v}{F^2} = \frac{v_{n,i}}{D_1^2 F_n^2} &= \frac{D_2 F_n^{2-\nu}(2c_1 F_n)^\nu}{D_1^2 F_n^2} \\ &\le \frac{D_2(2c_1)^\nu}{D_1^2} = \frac{64(2c_1)^{2-2\nu}(2c_1)^\nu}{D_2} \le 1\end{aligned}$$

and

$$\begin{aligned}M_{n,i} = M_{n,i}^{1-\nu} M_{n,i}^\nu &\le (2c_1 F_n)^{1-\nu} M_{n,i} \\ &\le \frac{8(2c_1)^{1-\nu} D_1 F_n^{2-\nu} M_{n,i}^\nu}{8 D_1 F_n} = \frac{v_{n,i}}{8 D_1 F_n}.\end{aligned}$$

Therefore, (C2) in Lemma 1.7.2 holds.

For (C3), we first note that

$$H_B(\delta, \mathcal{H}_{n,i}, \|\cdot\|_2) \le H_B(\delta/c_1, \mathcal{F}_{n,i}, \|\cdot\|_2) \le H_B(\delta/c_1, \mathcal{F}_n, \|\cdot\|_2),$$

where the first inequality follows from (A1) and the second inequality follows from $\mathcal{F}_{n,i} \subset \mathcal{F}_n$. Because $\int_{\zeta M_{n,i}/32}^{v_{n,i}^{1/2}} H_B(u, \mathcal{F}_n, \|\cdot\|_2)^{1/2} \mathrm{d}u / M_{n,i}$ is non-increasing in i,

$$\begin{aligned}&M_{n,i}^{-1} \int_{M_{n,i}/64}^{v_{n,i}^{1/2}} H_B^{1/2}(u, \mathcal{H}_{n,i}, \|\cdot\|_2)\mathrm{d}u \\ &\le M_{n,1}^{-1} \int_{M_{n,1}/64}^{v_{n,1}^{1/2}} H_B^{1/2}(u/c_1, \mathcal{F}_n, \|\cdot\|_2)\mathrm{d}u \\ &\le M_{n,1}^{-1} v_{n,1}^{1/2} H_B^{1/2}(M_{n,1}/(64c_1), \mathcal{F}_n, \|\cdot\|_2) \\ &\le (D_2 F_n^{2-\nu})^{1/2} M_{n,1}^{\nu/2-1} H_B^{1/2}(\epsilon_n^2/(128c_1), \mathcal{F}_n, \|\cdot\|_2) \\ &\le D_2^{1/2} F_n^{1-\nu/2} (\epsilon_n^2/2)^{\nu/2-1} H_B^{1/2}(\delta_n, \mathcal{F}_n, \|\cdot\|_2) \\ &\le c_3^{1/2} n^{1/2} D_2^{1/2} F_n^{1-\nu/2} (\epsilon_n^2/2)^{\nu/2-1} \left(\frac{\delta_n}{F_n}\right)^{1-\nu/2} \\ &= n^{1/2}(c_3 D_2)^{1/2} \left(\frac{2\delta_n}{\epsilon_n^2}\right)^{1-\nu/2} \le 2^{-23/2} n^{1/2},\end{aligned} \tag{1.7.6}$$

where the last inequality follows from that $c_3 D_2 = 2^{-11-6\nu} c_1^{\nu-2}$ and $\delta_n/\epsilon_n^2 \le$

$c_1/2^7$. Hence (C3) of Lemma 1.7.2 is satisfied. Furthermore, (1.7.6) implies that

$$\begin{aligned} H_B(v_{n,i}^{1/2}, \mathcal{H}_{n,i}, \|\cdot\|_2)^{1/2} &\leq \frac{M_{n,i}}{v_{n,i}^{1/2} - M_{n,i}/64} M_{n,i}^{-1} \int_{M_{n,i}/64}^{v_{n,i}^{1/2}} H_B^{1/2}(u, \mathcal{H}_{n,i}, \|\cdot\|_2) \mathrm{d}u \\ &\leq \frac{M_{n,i}}{v_{n,i}^{1/2} - M_{n,i}/64} n^{1/2} 2^{-23/2} \\ &\leq \frac{8}{7} \frac{M_{n,i}}{v_{n,i}^{1/2}} n^{1/2} 2^{-23/2} = \frac{1}{7 \times 2^{17/2}} \frac{M_{n,i}}{v_{n,i}^{1/2}} n^{1/2}, \end{aligned}$$

where the last inequality is due to that $v_{n,i}^{1/2} \geq M_{n,i}/8$. On the other hand, since $v_{n,i}/(8D_1F_n) \geq M_{n,i}$,

$$n \frac{M_{n,i}^2}{16(4v_{n,i} + M_{n,i}D_1F_n/3)} \geq n \frac{M_{n,i}^2}{(64 + 2/3)v_{n,i}}$$

which is larger than $\frac{1}{7^2 \times 2^{17}} \frac{M_{n,i}^2}{v_{n,i}} n$. Hence (C1) of Lemma 1.7.2 is met.

Applying Lemma 1.7.2 to each $\mathcal{H}_{n,i}$, (1.7.5) is further bounded as

$$\begin{aligned} \mathsf{P}\left\{\mathcal{E}_\phi(\hat{f}_n, f_\phi^*) \geq \epsilon_n^2\right\} &\leq \sum_{i=1}^{i_n^*} 3 \exp\left(-\frac{n M_{n,i}^2}{4(4v_{n,i} + M_{n,i}F_n/3)}\right) \\ &\leq \sum_{i=1}^{\infty} 3 \exp(-c_4 n M_{n,i}^2 / v_{n,i}\})) \\ &\leq \sum_{i=1}^{\infty} 3 \exp(-c_5 (2^i)^{2-\nu} n (\epsilon_n^2/F_n)^{2-\nu}) \\ &\leq c_6 \exp(-c_5 n (\epsilon_n^2/F_n)^{2-\nu}) \end{aligned}$$

for some positive constants c_4, c_5, and c_6, which leads to the desired result. □

1.7.3 Generic convergence rate for the hinge loss

We derive the convergence rate of the excess risk of the hinge loss under the conditions (A2), (A3), and (A5). Note that (A1) holds with $c_1 = 1$ for the

hinge loss. We adopt the following lemma for the variance bound (A4).

Lemma 1.7.4 (Lemma 6.1 of [85]). *Assume Assumption N with the noise exponent $q \in [0, \infty]$. Assume $\|f\|_\infty \le F$ for any $f \in \mathcal{F}$. For the hinge loss ϕ, we have that, for any $f \in \mathcal{F}$,*

$$\begin{aligned}&\mathsf{E}\left[\left(\phi(Yf(\mathbf{X})) - \phi(Yf_\phi^*(\mathbf{X}))\right)^2\right] \\ &\le c_{\eta,q}(F+1)^{(q+2)/(q+1)}\left(\mathsf{E}\left[\phi(Yf(\mathbf{X})) - \phi(Yf_\phi^*(\mathbf{X}))\right]\right)^{q/q+1},\end{aligned}$$

where $c_{\eta,q} = \left(\|(2\eta-1)^{-1}\|_{q,\infty}^q + 1\right)\mathbb{1}(q > 0) + 1$ and $\|(2\eta-1)^{-1}\|_{q,\infty}^q$ is defined by

$$\|(2\eta-1)^{-1}\|_{q,\infty}^q = \sup_{t>0}\left(t^q \mathsf{P}\left(\{\mathbf{X} : |(2\eta(\mathbf{X})-1)^{-1}| > t\}\right)\right).$$

Theorem 1.7.5. *Let ϕ be the hinge loss. Assume Assumption N with the noise exponent $q \in [0, \infty]$, and that (A2),(A3), and (A5) are met. Let $\epsilon_n^2 \asymp \max\{a_n, \delta_n\}$. Assume that $n(\epsilon_n^2/F_n)^{(q+2)/(q+1)} \gtrsim \log^{1+\kappa} n$ for an arbitrarily small constant $\kappa > 0$. Then the empirical ϕ-risk minimizer $\hat{f}_{\phi,n}$ over $\mathcal{F}_n$ satisfies*

$$\mathsf{E}\left[\mathcal{E}(\hat{f}_{\phi,n}, C^*)\right] \lesssim \epsilon_n^2, \tag{1.7.7}$$

where the expectation is taken over the training data.

Proof. By Zhang's inequality (Theorem 2.31 of [84]), we have

$$\mathcal{E}(\hat{f}_{\phi,n}, C^*) \le \mathcal{E}_\phi(\hat{f}_{\phi,n}, f_\phi^*).$$

Since (A4) is satisfied with $\nu = q/(q+1)$ by Lemma 1.7.4, Theorem 1.7.3 implies that

$$\mathsf{P}\left(\mathcal{E}(\hat{f}_{\phi,n}, C^*) \ge \epsilon_n^2\right) \lesssim \exp(-cn(\epsilon_n^2/F_n)^{(q+2)/(q+1)})$$

for some universal constant $c > 0$. Since $\mathcal{E}(\hat{f}_{\phi,n}, C^*)$ is bounded above by 1, the preceding display and the assumption $n(\epsilon_n^2/F_n)^{(q+2)/(q+1)} \gtrsim \log^{1+\kappa} n$ imply the desired result. □

1.7.4 Proof of Theorem 1.3.1

We first introduce the smooth function approximation result of DNNs.

Proposition 1.7.6. *For any function $f \in \mathcal{H}^{\beta,r}([0,1]^d)$ and any sufficiently small $\xi > 0$, there exists a neural network*

$$f^\circ \in \mathcal{F}^{\mathrm{DNN}}\left(L_0 \log(1/\xi), N_0 \xi^{-d/\beta}, S_0 \xi^{-d/\beta} \log(1/\xi), 1, F_0\right)$$

such that

$$\|f^\circ - f\|_\infty \le \xi, \tag{1.7.8}$$

where the constants L_0, N_0, S_0, and F_0 depend only on d, β and r.

Proof. Theorem 5 of [80] proves that for any $f \in \mathcal{H}^{\beta,r}([0,1]^d)$ and any integers $m \ge 1$ and $M \ge (\beta+1)^d \vee (r+1)e^d$, there exists a neural network $f^\circ \in \mathcal{F}^{\mathrm{DNN}}(L, N, S, 1, \infty)$ such that

$$\left\|f^\circ - f\right\|_\infty \le (2r+1)(1+d^2+\beta^2)6^d M 2^{-m} + r 3^\beta M^{-\beta/d},$$

where $L = 8 + (m+5)(1+\lceil \log_2(d \vee \beta) \rceil)$, $N = 6(d+\lceil\beta\rceil)M$, and $S = 141(d+\beta+1)^{3+d} M(m+6)$. By letting $M = (3^{-\beta}(2r)^{-1}\xi)^{-d/\beta}$ and

$$m = \log_2\left((2r+1)(1+d^2+\beta^2)6^d (3^{-\beta}(2r)^{-1}\xi)^{-d/\beta}(2/\xi)\right),$$

we have $L \lesssim \log(1/\xi)$, $N \lesssim \xi^{-d/\beta}$, $S \lesssim \xi^{-d/\beta}\log(1/\xi)$, and $\|f^\circ - f\|_\infty \le \xi$. Finally, because $\|f\|_\infty \le r$, we have $\|f^\circ\|_\infty \le r + \epsilon$, and hence we complete the proof with $F_0 \ge r + \xi$. □

Proof of Theorem 1.3.1. For a given ξ_n, by Proposition 1.7.6, there exists η_n° such that $\|\eta_n^\circ(\mathbf{x}) - \eta(\mathbf{x})\|_\infty \le \xi_n$ with the number of layers $\lesssim \log(1/\xi_n)$, the maximum number of hidden nodes $\lesssim \xi_n^{-d/\beta}$, sparsity $\lesssim \xi_n^{-d/\beta}\log(1/\xi_n)$, and the largest absolute value $\lesssim 1$. We construct the neural network f_n by

adding one layer to $\eta_n^\circ(\mathbf{x})$ as

$$f_n(\mathbf{x}) := 2\left[\sigma\left\{\frac{1}{\xi_n}\left(\eta_n^\circ(\mathbf{x})-\frac{1}{2}\right)\right\}-\sigma\left\{\frac{1}{\xi_n}\left(\eta_n^\circ(\mathbf{x})-\frac{1}{2}\right)-1\right\}\right]-1$$
$$=\begin{cases}1 & \text{if } \eta_n^\circ(\mathbf{x})\geq 1/2+\xi_n\\ 2(\eta_n^\circ(\mathbf{x})-1/2)/\xi_n-1 & \text{if } 1/2\leq\eta_n^\circ(\mathbf{x})<1/2+\xi_n\\ -1 & \text{if } \eta_n^\circ(\mathbf{x})<1/2.\end{cases}$$

We let

$$\mathcal{X}_{\eta,\xi} := \{\mathbf{x} : |2\eta(\mathbf{x})-1|>\xi\}$$

for $\xi>0$. Then, for $\mathbf{x}\in\mathcal{X}_{\eta,4\xi_n}$, $|f_n(\mathbf{x})-C^*(\mathbf{x})|=0$ because $\eta_n^\circ(\mathbf{x})-1/2=(\eta(\mathbf{x})-1/2)-(\eta_n^\circ(\mathbf{x})-\eta(\mathbf{x}))\geq\xi_n$ when $2\eta(\mathbf{x})-1>4\xi_n$ and $\eta_n^\circ(\mathbf{x})-1/2<-\xi_n$ when $2\eta(\mathbf{x})-1<4\xi_n$. Therefore, Assumption N implies

$$\begin{aligned}\mathsf{E}[\phi(Yf_n(\mathbf{X}))-\phi(YC^*(\mathbf{X}))] &= \int |f_n(\mathbf{x})-C^*(\mathbf{x})||2\eta(\mathbf{x})-1|\mathrm{d}P_X(\mathbf{x})\\ &= \int_{[0,1]^d\setminus\mathcal{X}_{\eta,\xi}} |f_n(\mathbf{x})-C^*(\mathbf{x})||2\eta(\mathbf{x})-1|\mathrm{d}P_X(\mathbf{x})\\ &\leq 8\xi_n P(\{\mathbf{X}:|2\eta(\mathbf{X})-1|\leq 4\xi_n\}) \asymp \xi_n^{q+1},\end{aligned}$$

where the first equality follows from Theorem 2.31 of [84] and the inequality in the last line holds since $\|f_n(\mathbf{x})\|_\infty\leq 1$.

Note that $f_n\in\mathcal{F}^{\mathrm{DNN}}(L_n,N_n,S_n,B_n,1)$ with $L_n\lesssim\log(1/\xi_n)$, $N_n\lesssim\xi_n^{-d/\beta}$, $S_n\lesssim\xi_n^{-d/\beta}\log(1/\xi_n)$, and $B_n\lesssim\xi_n^{-1}$. If we take $\epsilon_n^2=\xi_n^{q+1}$, by Proposition 1.7.1, we obtain

$$\begin{aligned}&\log\mathcal{N}(\epsilon_n^2,\mathcal{F}^{\mathrm{DNN}}(L_n,N_n,S_n,B_n,1),\|\cdot\|_\infty)\\ &\leq 2L_n(S_n+1)\log\left((\epsilon_n^2)^{-1}(L_n+1)(N_n+1)(B_n\vee 1)\right)\\ &\lesssim \log(\xi_n^{-1})\xi_n^{-d/\beta}\log(\xi_n^{-1})\log(\epsilon_n^{-1})\\ &\lesssim (\epsilon_n^2)^{-d/(q+1)\beta}\log^3(\epsilon_n^{-1}).\end{aligned}$$

Due to the inequality in (1.7.1), (A5) is satisfied if we choose ϵ_n satisfying

$$(\epsilon_n^2)^{\frac{q+2}{q+1}+\frac{d}{\beta(q+1)}}\gtrsim n^{-1}\log^3(\epsilon_n^{-1}),$$

which leads to the best possible convergence rate

$$\epsilon_n^2 = \left(\frac{\log^3 n}{n}\right)^{\frac{\beta(q+1)}{\beta(q+2)+d}}$$

and completes the proof based on Theorem 1.7.5. □

1.7.5 Proof of Theorem 1.3.2

The following proposition given by [76] proves that DNNs are good at approximating piecewise constant functions with smooth boundaries.

Proposition 1.7.7 (Corollary 3.7 of [76]). *Let $d \geq 2$, $\alpha, r > 0, K \in \mathbb{N}$, and $T \in \mathbb{N}$. For any $C \in \mathcal{C}^{\alpha,r,K,T}$ and any sufficiently small $\xi > 0$, there exists a neural network*

$$f^\circ \in \mathcal{F}^{\mathrm{DNN}}\left(L_0 \log\left(1/\xi\right), N_0 \xi^{-(d-1)/\alpha}, S_0 \xi^{-(d-1)/\alpha} \log\left(1/\xi\right), B_0 \xi^{-1}, 1\right),$$

where the positive constants L_0, N_0, S_0 and B_0 depend only on d, α, r, K, and T, such that

$$\left\|f^\circ - C\right\|_1 \leq \xi.$$

Lemma 1.7.8. *Assume that P satisfies Assumption D. Let ϕ is the hinge loss. Let $\{\xi_n\}_{n\in\mathbb{N}}$ be a positive sequence such that $\xi_n \downarrow 0$, and let $L_n \lesssim \log(1/\xi_n)$, $N_n \lesssim \xi_n^{-(d-1)/\alpha}$, $S_n \lesssim \xi_n^{-(d-1)/\alpha} \log(1/\xi_n)$ and $B_n \lesssim \xi_n^{-1}$. Then, there exists $f_n \in \mathcal{F}^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, 1)$ such that*

$$\mathcal{E}_\phi(f_n, f_\phi^*) \lesssim \xi_n.$$

Proof. By Proposition 1.7.7, there exists $f_n \in \mathcal{F}^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, 1)$ such that $\|f_n - C^*\| \leq \xi_n$. Since ϕ is Lipschitz with constant 1, Assumption D

implies

$$\begin{aligned}\mathcal{E}_\phi(f_n, f_\phi^*) &= \mathsf{E}[\phi(Yf_n(\mathbf{X})) - \phi(YC^*(\mathbf{X}))] \\ &\le \mathsf{E}|Yf_n(\mathbf{X}) - YC^*(\mathbf{X})| \\ &\lesssim \mathsf{E}|f_n(\mathbf{X}) - C^*(\mathbf{X})| \\ &\lesssim \|f_n - C^*\|_1 \lesssim \xi_n,\end{aligned}$$

which is the desired result. □

Proof of Theorem 1.3.2. We will check the conditions (A2), (A3), and (A5) in Section 1.7.2, and apply Theorem 1.7.5 to complete the proof.

For given ξ_n, let $L_n \lesssim \log(1/\xi_n)$, $N_n \lesssim \xi_n^{-(d-1)/\alpha}$, $S_n \lesssim \xi_n^{-(d-1)/\alpha}\log(1/\xi_n)$ and $B_n \lesssim \xi_n^{-1}$. Then, (A2) and (A3) hold with $\mathcal{F}_n = \mathcal{F}^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, 1)$, $a_n = \xi_n$ and $F_n = 1$ due to Lemma 1.7.8.

Let $\epsilon_n^2 = \xi_n$. Then, by Proposition 1.7.1,

$$\begin{aligned}&\log \mathcal{N}(\epsilon_n^2, \mathcal{F}^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, 1), \|\cdot\|_\infty) \\ &\le 2L_n(S_n+1)\log\left((\epsilon_n^2)^{-1}(L_n+1)(N_n+1)(B_n \vee 1)\right) \\ &\lesssim \xi_n^{-(d-1)/\alpha}\log^2(\xi_n^{-1})\log(\epsilon_n^{-1}) \\ &\lesssim \epsilon_n^{-2(d-1)/\alpha}\log^3(\epsilon_n^{-1}).\end{aligned}$$

Due to the ineqaulity in (1.7.1), (A5) is satisfied if we choose ϵ_n satisfying

$$(\epsilon_n^2)^{\frac{q+2}{q+1}+\frac{(d-1)}{\alpha}} \gtrsim n^{-1}\log^3(\epsilon_n^{-1}),$$

which leads to the best possible convergence rate

$$\epsilon_n^2 = \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha(q+1)}{\alpha(q+2)+(d-1)(q+1)}}$$

and completes the proof by Theorem 1.7.5. □

1.7.6 Proof of Theorem 1.3.3

Proof of Theorem 1.3.3. The main idea of the proof is to find f_n whose approximation error is smaller than that in Theorem 1.3.2. For a given $\xi_n > 0$, by Proposition 1.7.6, we have that for any $t \in [T]$ and $k \in [K]$, there is a neural network $\tilde{g}^\circ_{n,t,k} \in \mathcal{F}^{\mathrm{DNN}}(L'_n, N'_n, S'_n, 1, F_0)$ for some $F_0 > 0$ such that $\|\tilde{g}^\circ_{n,t,k} - g^*_{t,k}\|_\infty \le \xi_n/2$ with $L_n \lesssim \log(1/\xi_n)$, $N_n \lesssim \xi_n^{-(d-1)/\alpha}$ and $S_n \lesssim \xi_n^{-(d-1)/\alpha}\log(1/\xi_n)$. Let $g^\circ_{n,t,k} = \tilde{g}^\circ_{n,t,k} + \xi_n/2$. Then $g^\circ_{n,t,k}$ is a neural network such that $\|g^\circ_{n,t,k} - g^*_{t,k}\|_\infty \le \xi_n$ and $g^\circ_{n,t,k}(\mathbf{x}) \ge g^*_{t,k}(\mathbf{x})$ for any $\mathbf{x} \in [0,1]^{d-1}$. We construct the neural network $f_{n,t,k}$ as

$$
\begin{aligned}
&f_{n,t,k}(\mathbf{x}) \\
&:= 2\left[\sigma\left\{\frac{1}{\xi_n}\left(x_{j^*_{t,k}} - g^\circ_{n,t,k}\left(\mathbf{x}_{-j^*_{t,k}}\right)\right)\right\} - \sigma\left\{\frac{1}{\xi_n}\left(x_{j^*_{t,k}} - g^\circ_{n,t,k}\left(\mathbf{x}_{-j^*_{t,k}}\right)\right) - 1\right\}\right] - 1 \\
&= \begin{cases} 1 & \text{if } x_{j^*_{t,k}} \ge g^\circ_{n,t,k}(\mathbf{x}_{-j^*_{t,k}}) + \xi_n \\ 2\xi_n^{-1}\left(x_{j^*_{t,k}} - g^\circ_{n,t,k}(\mathbf{x}_{-j^*_{t,k}})\right) - 1 & \text{if } g^\circ_{n,t,k}(\mathbf{x}_{-j^*_{t,k}}) \le x_{j^*_{t,k}} < g^\circ_{n,t,k}(\mathbf{x}_{-j^*_{t,k}}) + \xi_n \\ -1 & \text{if } x_{j^*_{t,k}} < g^\circ_{n,t,k}(\mathbf{x}_{-j^*_{t,k}}). \end{cases}
\end{aligned}
$$

Note that $\|f_{n,t,k}\|_\infty \le 1$.

We then construct the neural network f_n as

$$
f_n(\mathbf{x}) := \sum_{t=1}^{T} \sigma\left(\sum_{k=1}^{K} f_{n,t,k}(\mathbf{x}) - (K-2)\right) - 1.
$$

We first show that $f_n(\mathbf{x}) \le 1$ for any $\mathbf{x} \in [0,1]^d$. For this, we introduce the notation

$$
A^\circ_t := \left\{\mathbf{x} \in [0,1]^d : x_{j^*_{t,k}} \ge g^\circ_{n,t,k}\left(\mathbf{x}_{-j^*_{t,k}}\right), \forall k \in [K]\right\}. \tag{1.7.9}
$$

Then we have

$$\begin{aligned} f_n(\mathbf{x}) &\leq 2\sum_{t=1}^{T}\mathbb{1}\left(\sum_{k=1}^{K} f_{n,t,k}(\mathbf{x}) > K-2\right) - 1 \\ &\leq 2\sum_{t=1}^{T}\mathbb{1}\left(\mathbf{x}\in A_t^\circ\right) - 1, \end{aligned}$$

where the second inequality is due to that for $\mathbf{x} \in (A_t^\circ)^c$ there exists $k \in [K]$ such that $f_{n,t,k}(\mathbf{x}) = -1$, and thus $\sum_{k=1}^{K} f_{n,t,k}(\mathbf{x}) \leq K-2$. Let

$$A_t^* := \left\{\mathbf{x}\in[0,1]^d : x_{j_{t,k}^*} \geq g_{t,k}^*\left(\mathbf{x}_{-j_{t,k}^*}\right), \forall k \in [K]\right\}. \tag{1.7.10}$$

Since $g_{n,t,k}^\circ(\mathbf{x}_{-j_{t,k}^*}) \geq g_{t,k}^*(\mathbf{x}_{-j_{t,k}^*})$ for any $t \in [T]$ and any $k \in [K]$, it follows that $A_t^\circ \subset A_t^*$ for any $t \in [T]$. Since $(A_t^*)_{t\in[T]}$ are disjoint, so are $(A_t^\circ)_{t\in[T]}$. Hence $2\sum_{t=1}^{T}\mathbb{1}\left(\mathbf{x}\in A_t^\circ\right) - 1 \leq 1$ for any $\mathbf{x} \in [0,1]^d$ and we conclude that $f_n(\mathbf{x}) \leq 1$ for any $\mathbf{x} \in [0,1]^d$.

For the lower bound of f_n, note that

$$f_{n,t,k}(\mathbf{x}) \geq 2\mathbb{1}\left(x_{j_{t,k}^*} \geq g_{n,t,k}^\circ\left(\mathbf{x}_{-j_{t,k}^*}\right) + \xi_n\right) - 1,$$

and thus

$$\begin{aligned} f_n(\mathbf{x}) &\geq \sum_{t=1}^{T}\sigma\left(2\sum_{k=1}^{K}\mathbb{1}\left(x_{j_{t,k}^*} \geq g_{n,t,k}^\circ\left(\mathbf{x}_{-j_{t,k}^*}\right) + \xi_n\right) - 2(K-1)\right) - 1 \\ &= 2\sum_{t=1}^{T}\mathbb{1}\left(x_{j_{t,k}^*} \geq g_{n,t,k}^\circ\left(\mathbf{x}_{-j_{t,k}^*}\right) + \xi_n, \forall k \in [k]\right) - 1 \\ &= 2\sum_{t=1}^{T}\mathbb{1}\left(\mathbf{x}\in A_{t,\xi_n}^\circ\right) - 1 := C_n^\circ(\mathbf{x}). \end{aligned} \tag{1.7.11}$$

where A_{t,ξ_n}° is defined as

$$A_{t,\xi_n}^\circ := \left\{\mathbf{x}\in[0,1]^d : x_{j_{t,k}^*} \geq g_{t,k}^*\left(\mathbf{x}_{-j_{t,k}^*}\right) + \xi_n, \forall k \in [K]\right\}.$$

Since $A^\circ_{t,\xi_n} \subset A^\circ_t \subset A^*_t$ for any $t \in [T]$, $(A^\circ_{t,\xi_n})_{t\in[T]}$ are disjoint and hence $C^\circ_n(\mathbf{x}) \in \{-1, 1\}$ for any $\mathbf{x} \in [0,1]^d$.

Recall that $g^\circ_{n,t,k}(\mathbf{x}_{-j^*_{t,k}}) \geq g^*_{t,k}(\mathbf{x}_{-j^*_{t,k}})$ for any $t \in [T]$ and any $k \in [K]$. Thus for an input $\mathbf{x} \in [0,1]^d$ such that $C^*(\mathbf{x}) = -1$, there is $k_t \in [K]$ such that $f_{n,t,k_t}(\mathbf{x}) = -1$ for all $t \in [T]$, and hence $\sum_{k=1}^K f_{n,t,k}(\mathbf{x}) \leq K - 2$. Thus, we conclude that $f_n(\mathbf{x}) = -1$ whenever $C^*(\mathbf{x}) = -1$, which implies

$$\begin{aligned}
|f_n(\mathbf{x}) - C^*(\mathbf{x})| &= |f_n(\mathbf{x}) - C^*(\mathbf{x})| \mathbb{1}\{C^*(\mathbf{x}) = 1\} \\
&\leq \left|C^\circ_n(\mathbf{x}) - C^*(\mathbf{x})\right| \mathbb{1}\{C^*(\mathbf{x}) = 1\} \\
&= 2\mathbb{1}\{C^\circ_n(\mathbf{x}) \neq C^*(\mathbf{x})\}\mathbb{1}\{C^*(\mathbf{x}) = 1\} \\
&\leq 2\mathbb{1}\{C^\circ_n(\mathbf{x}) \neq C^*(\mathbf{x})\},
\end{aligned}$$

where the first inequality follows from (1.7.11) and the fact that $f_n(\mathbf{x}) \leq 1$. Since $\left\|g^\circ_{n,t,k} + \xi_n - g^*_{t,k}\right\|_\infty \leq 2\xi_n$ for any $t \in [T]$ and any $k \in [K]$, the condition (R) implies

$$\begin{aligned}
\mathcal{E}_\phi(f_n, C^*) &= \int |f_n(\mathbf{x}) - C^*(\mathbf{x})||2\eta(\mathbf{x}) - 1| \mathrm{d}\mathsf{P}_X(\mathbf{x}) \\
&\leq 2\int \mathbb{1}\{C^\circ_n(\mathbf{x}) \neq C^*(\mathbf{x})\}|2\eta(\mathbf{x}) - 1| \mathrm{d}\mathsf{P}_X(\mathbf{x}) \\
&= 2\{\mathcal{E}(C^\circ_n) - \mathcal{E}(C^*)\} \\
&\leq 2^{1+(q+1)/q} c_{\mathrm{R}} \xi_n^{(q+1)/q}.
\end{aligned}$$

On the other hand, $f_n \in \mathcal{F}^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, 1)$ such that with $L_n \lesssim \log(1/\xi_n)$, $N_n \lesssim \xi_n^{-(d-1)/\alpha}$, $S_n \lesssim \xi_n^{-(d-1)/\alpha}\log(1/\xi_n)$ and $B_n \lesssim \xi_n^{-1}$. By taking $\epsilon_n^2 = \xi_n^{(q+1)/q}$, we have

$$\log \mathcal{N}(\epsilon_n^2, \mathcal{F}^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, 1), \|\cdot\|_\infty) \lesssim (\epsilon_n^2)^{-q(d-1)/\alpha(q+1)} \log^3(\epsilon_n^{-1}).$$

(A5) is satisfied if we choose ϵ_n satisfying

$$(\epsilon_n^2)^{\frac{q+2}{q+1} + \frac{q(d-1)}{\alpha(q+1)}} \gtrsim n^{-1}\log^3(\epsilon_n^{-1}),$$

which leads to the best possible convergence rate

$$\epsilon_n^2 = \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha(q+1)}{\alpha(q+2)+(d-1)q}}$$

and completes the proof by Theorem 1.7.5. □

1.7.7 Proof of Theorem 1.3.4

The main technique of the proof is to approximate a piecewise constant function using a DNN with respect to the supremum norm on a specific subset of the domain, where this subset depends on the function to be approximated.

Let $d \geq 2$, $\alpha, r > 0$, and $K \in \mathbb{N}$. Let $A_1, \ldots, A_T \in \mathcal{A}^{\alpha,r,K}$ be disjoint with the form

$$A_t := \bigcap_{k=1}^{K} \left\{\mathbf{x} \in [0,1]^d : x_{j_{t,k}} - g_{t,k}(\mathbf{x}_{-j_{t,k}}) \geq 0\right\}. \tag{1.7.12}$$

Let $T \in \mathbb{N}$, and let

$$C(\mathbf{x}) := 2\sum_{t=1}^{T} \mathbb{1}(\mathbf{x} \in A_t) - 1. \tag{1.7.13}$$

Let

$$A_t^c := [0,1]^d \setminus A_t = \bigcup_{k=1}^{K} \left\{\mathbf{x} \in [0,1]^d : x_{j_{t,k}} - g_{t,k}(\mathbf{x}_{-j_{t,k}}) < 0\right\} \tag{1.7.14}$$

and

$$A_{t,\xi} := \left\{\mathbf{x} \in [0,1]^d : x_{j_{t,k}} - g_{t,k}(\mathbf{x}_{-j_{t,k}}) > \xi, \forall k \in [K]\right\}. \tag{1.7.15}$$

for a given $\xi > 0$. Define a subset $\mathcal{X}_{A_{1:T},\xi}$ of $[0,1]^d$ such that

$$\mathcal{X}_{A_{1:T},\xi} := \bigcap_{t=1}^{T} \left(A_t^c \cup A_{t,\xi}\right). \tag{1.7.16}$$

The following theorem proves that a DNN recovers $C(\mathbf{x})$ exactly on $\mathcal{X}_{A_{1:T},\xi}$.

Proposition 1.7.9. *Let $d \geq 2$, $\alpha, r > 0, K \in \mathbb{N}$, and $T \in \mathbb{N}$. For any $C \in \mathcal{C}^{\alpha,r,K,T}$ and a sufficiently small $\xi > 0$, there exists a neural network*

$$f^\circ \in \mathcal{F}^{\text{DNN}}\left(L_0 \log\left(1/\xi\right), N_0 \xi^{-(d-1)/\alpha}, S_0 \xi^{-(d-1)/\alpha} \log\left(1/\xi\right), B_0 \xi^{-1}, 1\right),$$

where the positive constants L_0, N_0, S_0, and B_0 depend only on d, α, r, K and T, such that

$$\sup_{\mathbf{x} \in \mathcal{X}_{A_{1:T},\xi}} \left|f^\circ(\mathbf{x}) - C(\mathbf{x})\right| = 0.$$

Proof. The proof is deferred to Section 1.7.10. □

Proof of Theorem 1.3.4. Let $\{\xi_n\}_{n \in \mathbb{N}}$ be a positive sequence such that $\xi_n \downarrow 0$. By Proposition 1.7.9, there exists $f_n \in \mathcal{F}_n = \mathcal{F}^{\text{DNN}}(L_n, N_n, S_n, B_n, 1)$ such that

$$\sup_{\mathbf{x} \in \mathcal{X}_{A_{1:T},\xi_n}} \left|f_n(\mathbf{x}) - C(\mathbf{x})\right| = 0$$

with $L_n \lesssim \log(1/\xi_n)$, $N_n \lesssim \xi_n^{-(d-1)/\alpha}$, $S_n \lesssim \xi_n^{-(d-1)/\alpha} \log(1/\xi_n)$, and $B_n \lesssim \xi_n^{-1}$.

We will show that

$$\mathcal{X}^*_\xi := \{\mathbf{x} : \text{dist}(\mathbf{x}, D^*) > \xi\} \subset \mathcal{X}_{A_{1:T},\xi}, \tag{1.7.17}$$

for any $\xi > 0$. Suppose that $\mathbf{x} \in [0,1]^d \setminus \mathcal{X}_{A_{1:T},\xi}$. Then there are $t \in [T]$ and $k \in [K]$ such that $|x_{j_{t,k}} - g_{t,k}(\mathbf{x}_{-j_{t,k}})| \leq \xi$. Let $\mathbf{x}^*$ be the d-dimensional vector where the $j_{t,k}$-th component is equal to $g_{t,k}(\mathbf{x}_{-j_{t,k}})$ and the other components are the same as the corresponding components of $\mathbf{x}$, i.e., $x^*_{j_{t,k}} = g_{t,k}(\mathbf{x}_{-j_{t,k}})$ and $\mathbf{x}^*_{-j_{t,k}} = \mathbf{x}_{-j_{t,k}}$. Clearly, $\mathbf{x}^*$ is on the decision boundary D^*. Since $\|\mathbf{x} - \mathbf{x}^*\|_2 = |x_{j_{t,k}} - g_{t,k}(\mathbf{x}_{-j_{t,k}})| \leq \xi$, it follows that $\text{dist}(\mathbf{x}, D^*) \leq \xi$, and hence, $\mathbf{x} \in [0,1]^d \setminus \mathcal{X}^*_\xi$, which proves (1.7.17).

Since $f_n(\mathbf{x}) - C^*(\mathbf{x}) = 0$ for any $\mathbf{x} \in \mathcal{X}_{A_{1:T},\xi_n}$, (1.7.17) and Assumption M imply

$$\begin{aligned}
&\mathsf{E}[\phi(Yf_n(\mathbf{X})) - \phi(YC^*(\mathbf{X}))] \\
&= \int |f_n(\mathbf{x}) - C^*(\mathbf{x})||2\eta(\mathbf{x}) - 1| \mathrm{dP}_X(\mathbf{x}) \\
&= \int_{[0,1]^d \setminus \mathcal{X}_{A_{1:T},\xi_n}} |f_n(\mathbf{x}) - C^*(\mathbf{x})||2\eta(\mathbf{x}) - 1| \mathrm{dP}_X(\mathbf{x}) \\
&\le 2\mathsf{P}\left(\left\{\mathbf{X} : \mathbf{X} \in [0,1]^d \setminus \mathcal{X}_{A_{1:T},\xi_n}\right\}\right) \\
&\le 2\mathsf{P}\left(\left\{\mathbf{X} : \mathbf{X} \in [0,1]^d \setminus \mathcal{X}^*_{\xi_n}\right\}\right) \le 2c_{\mathrm{M}}\xi_n^{\gamma}.
\end{aligned}$$

By taking $\epsilon_n^2 = \xi_n^{\gamma}$, it follows that

$$\log \mathcal{N}(\epsilon_n^2, \mathcal{F}^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, 1), \|\cdot\|_\infty) \lesssim (\epsilon_n^2)^{-(d-1)/\alpha\gamma} \log^3(\epsilon_n^{-1}).$$

Hence, (A5) is satisfied if we choose ϵ_n satisfying

$$(\epsilon_n^2)^{\frac{q+2}{q+1} + \frac{d-1}{\alpha\gamma}} \gtrsim n^{-1} \log^3(\epsilon_n^{-1}),$$

which leads to the best possible convergence rate

$$\epsilon_n^2 = \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha(q+1)}{\alpha(q+2)+(d-1)(q+1)/\gamma}},$$

and completes the proof by Theorem 1.7.5. □

1.7.8 Proof of Theorem 1.4.1

In this section, we write $\phi(f) = \phi(Yf(\mathbf{X}))$ for any $f \in \mathcal{F}$ for simplicity. We need the following three lemmas. The first two lemmas are the restatements of Theorem 4.3 and Lemma 6.1 of [11].

Lemma 1.7.10 (Theorem 4.3 of [11]). *Let ϕ be a loss function and assume that there exists $f^* \in \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{E}_\phi(f)$. Let $(\mathcal{F}_m)_{m\in\mathcal{M}}$, $\mathcal{F}_m \subset \mathcal{F}$ be a countable collection of classes of functions and assume there exist the followings:*

- *a pseudo-distance ρ on $\mathcal{F}$;*
- *a sequence of sub-root functions $(\psi_m)_{m\in\mathcal{M}}$*
- *two positive sequences $(b_m)_{m\in\mathcal{M}}$ and $(c_m)_{m\in\mathcal{M}}$;*

such that

(H1) $\|\phi(f)\|_\infty \leq b_m$, *for any* $m \in \mathcal{M}$ *and any* $f \in \mathcal{F}_m$;

(H2) $\mathsf{Var}(\phi(f) - \phi(f')) \leq \rho^2(f, f')$ *for any* $f, f' \in \mathcal{F}$;

(H3) $\rho^2(f, f^*) \leq c_m \mathcal{E}_\phi(f, f^*)$ *for any* $m \in \mathcal{M}$ *and any* $f \in \mathcal{F}_m$;

(H4) if u_m^* *denotes the solution of* $\psi_m(u) = u/c_m$, *for any* $m \in \mathcal{M}$, *any* $f_0 \in \mathcal{F}_m$, *and any* $u \geq u_m^*$,

$$\mathsf{E}\left[\sup_{f\in\mathcal{F}_m, \rho^2(f,f_0)\leq u} \left\{(\mathcal{E}_\phi(f) - \mathcal{E}_\phi(f_0)) - (\mathcal{E}_{\phi,n}(f) - \mathcal{E}_{\phi,n}(f_0))\right\}\right] \leq \psi_m(u).$$

Let $(z_m)_{m\in\mathcal{M}}$ be a sequence of real numbers such that $\sum_{m\in\mathcal{M}} \mathrm{e}^{-z_m} \leq 1$. We assume that families $(b_m)_{m\in\mathcal{M}}$, $(c_m)_{m\in\mathcal{M}}$ $(z_m)_{m\in\mathcal{M}}$ are ordered the same way by which we mean that $z_m < z_{m'}$ implies $b_m \leq b_{m'}$ and $c_m \leq c_{m'}$ for any $m, m' \in \mathcal{M}$. Let $K > 1$ be some real number to be fixed in advance. Let $\mathrm{pen}_n(m)$ be a penalty function such that, for each $m \in \mathcal{M}$

$$\mathrm{pen}_n(m) \geq 250K\frac{u_m^*}{c_m} + \frac{(75Kc_m + 28b_m)\left(z_m + \log 2 + 3\log(75Kc_m + 28b_m)\right)}{3n} \tag{1.7.18}$$

Let $\hat{f}_m = \mathrm{argmin}_{f\in\mathcal{F}_m} \mathcal{E}_{\phi,n}(f)$ for each $m \in \mathcal{M}$ and let

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \left[\mathcal{E}_{\phi,n}(\hat{f}_m) + \mathrm{pen}_n(m)\right].$$

Then the following inequality holds:

$$\mathsf{E}\left[\mathcal{E}_\phi\left(\hat{f}_{\hat{m}}, f^*\right)\right] \leq \frac{K + 1/5}{K - 1} \inf_{m\in\mathcal{M}} \left(\inf_{f\in\mathcal{F}_m} \mathcal{E}_\phi\left(f, f^*\right) + 2\mathrm{pen}_n(m)\right). \tag{1.7.19}$$

Lemma 1.7.11 (Lemma 6.10 of [11]). *Let $\mathcal{G}$ be a class of real-valued functions which is separable in the supremum norm, containing the null function, and such that every $g \in \mathcal{G}$ satisfies $\|g\|_\infty \leq G$ and $\mathsf{E}(g^2) \leq u$. Then*

$$\mathsf{E}\left[\sup_{g\in\mathcal{G}}\left|\mathsf{E}[g(Z)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right|\right] \leq \frac{24}{\sqrt{n}}\int_0^{\sqrt{u}} \sqrt{\log \mathcal{N}(u', \mathcal{G}, \|\cdot\|_\infty)}\mathrm{d}u' + \frac{G\log\mathcal{N}(\sqrt{u}, \mathcal{G}, \|\cdot\|_\infty)}{n}.$$

Lemma 1.7.12. *For any $m > 1$, we have*

$$\int_0^x \log^{1/2}\frac{m}{\epsilon}\mathrm{d}\epsilon \leq 2x\log^{1/2}\frac{em}{x},$$

for every $x \in (0, m)$.

Proof. Both the left-hand and the right-hand terms are equal at $x = 0$. But the derivates of the function on the left-hand side is smaller that of the function on right-hand side. □

Proof of Theorem 1.4.1. The result is a direct consequence of Lemma 1.7.10 with $m = \omega$ and $\mathcal{M} = \mathcal{A}_n$. We will check the conditions (H1)-(H4) in Lemma 1.7.10. Since the hinge loss is Lipschitz with constant 1, (H1) is met with $b_\omega = 1$. We set $\rho^2(f, f') = \mathsf{E}(\phi(f) - \phi(f'))^2$. Then (H2) is trivially satisfied and (H3) is satisfied with $c_\omega = 1$ by Lemma 1.7.4. For (H4) we apply Lemma 1.7.11 to the class $\mathcal{G}_\omega := \{\phi(f) - \phi(f_0) : f \in \mathcal{F}_{n,\omega}^{\mathrm{DNN}}\}$ for $f_0 \in \mathcal{F}_{n,\omega}^{\mathrm{DNN}}$. Then due to Proposition 1.7.1, Lemma 1.7.11 and Lemma 1.7.12, and the fact that ϕ is Lipschitz with constant 1, we have that

$$\mathsf{E}\left[\sup_{f\in\mathcal{F}_{n,\omega}^{\mathrm{DNN}}, \rho^2(f,f_0)\leq u}\Big[\{\mathcal{E}_\phi(f) - \mathcal{E}_\phi(f_0)\} - \{\mathcal{E}_{\phi,n}(f) - \mathcal{E}_{\phi,n}(f_0)\}\Big]\right] \leq \psi_{n,\omega}(u),$$

where

$$\psi_{n,\omega}(u) := 24\sqrt{2u}\sqrt{\frac{(S_{n,\omega}+1)L_{n,\omega}}{n}}\log^{1/2}\left(\frac{e(L_{n,\omega}+1)(N_{n,\omega}+1)(B_{n,\omega}\vee 1)}{\sqrt{u}}\right) + 4\frac{(S_{n,\omega}+1)L_{n,\omega}}{n}\log\left(\frac{(L_{n,\omega}+1)(N_{n,\omega}+1)(B_{n,\omega}\vee 1)}{\sqrt{u}}\right).$$

Let

$$u^{\#}_{n,\omega} := 1250\frac{(S_{n,\omega}+1)L_{n,\omega}}{n}\log\left(n(N_{n,\omega}+1)(B_{n,\omega}\vee 1)\right).$$

Note that $L_{n,\omega} \lesssim \log(1/\xi_{n,\omega})$, $N_{n,\omega} \lesssim \xi_{n,\omega}^{-(d-1)/\omega}$, $S_{n,\omega} \lesssim \xi_{n,\omega}^{-(d-1)/\omega}\log(1/\xi_{n,\omega})$ and $B_n \lesssim \xi_{n,\omega}^{-1}$, and so we have $u^{\#}_{n,\omega} \geq e^2(L_{n,\omega}+1)^2/n^2$ for sufficiently large n. Thus we have

$$\begin{aligned}
\psi_{n,\omega}(u^{\#}_{n,\omega}) \leq & 24\sqrt{2}(25\sqrt{2})\frac{(S_{n,\omega}+1)L_{n,\omega}}{n}\log^{1/2}\left(n(N_{n,\omega}+1)(B_{n,\omega}\vee 1)\right) \\
& \times \log^{1/2}\left(\frac{e(L_{n,\omega}+1)(N_{n,\omega}+1)(B_{n,\omega}\vee 1)}{\sqrt{e^2(L_{n,\omega}+1)^2/n^2}}\right) \\
& + 4\frac{(S_{n,\omega}+1)L_{n,\omega}}{n}\log\left(\frac{(L_{n,\omega}+1)(N_{n,\omega}+1)(B_{n,\omega}\vee 1)}{\sqrt{e^2(L_{n,\omega}+1)^2/n^2}}\right). \\
\leq & 24\sqrt{2}(25\sqrt{2})\frac{(S_{n,\omega}+1)L_{n,\omega}}{n}\log\left(n(N_{n,\omega}+1)(B_{n,\omega}\vee 1)\right) \\
& + 4\frac{(S_{n,\omega}+1)L_{n,\omega}}{n}\log\left(n(N_{n,\omega}+1)(B_{n,\omega}\vee 1)\right) \\
\leq & 1204\frac{(S_{n,\omega}+1)L_{n,\omega}}{n}\log\left(n(N_{n,\omega}+1)(B_{n,\omega}\vee 1)\right) \\
\leq & (25\sqrt{2})^2\frac{(S_{n,\omega}+1)L_{n,\omega}}{n}\log\left(n(N_{n,\omega}+1)(B_{n,\omega}\vee 1)\right) = u^{\#}_{n,\omega}.
\end{aligned}$$

Since $\psi_{n,\omega}(u)$ is a sub-root function, we have that $u^{*}_{n,\omega} \leq u^{\#}_{n,\omega}$, where $u^{*}_{n,\omega}$ is the solution of the equation $\psi_{n,\omega}(u) = u$. If we replace u^{*}_{m} in (1.7.18) by $u^{\#}_{n,\omega}$, we can see that the penaltiy function (1.4.2) satisfies (1.7.18) with $K = 2$ and hence (1.7.19) holds with $K = 2$.

Now we are ready to prove (1.4.3). Let

$$A_{n,\omega} := \xi_{n,\omega}^{-\frac{d-1}{\omega}} = \left(\frac{n}{\log^3 n}\right)^{\frac{d-1}{\omega+d-1}}.$$

By Lemma 1.7.8, we have that for any $C^* \in \mathcal{C}^{\alpha,r,K,T}$ there exists a neural network $f_{n,\omega} \in \mathcal{F}_{n,\omega}^{\mathrm{DNN}}$ such that $\mathcal{E}_\phi(f_{n,\omega}, C^*) \lesssim A_{n,\omega}^{-\alpha/(d-1)}$. Also, it is true that $\mathrm{pen}_n(\omega) \lesssim A_{n,\omega} \log^3 n/n$.

Let $k_{n,\alpha} \in \{1, \dots, \lfloor \log n \rfloor\}$ be the integer such that $k_{n,\alpha}/\log n \le \alpha/(\alpha + d - 1) < (k_{n,\alpha}+1)/\log n$. By the assumption that $\alpha \in (1/\tau, \tau)$, we have $w_{n,\alpha} := \frac{d-1}{\log n/k_{n,\alpha}-1} \in \mathcal{A}_n$. Then (1.7.19) with $K = 2$ implies that

$$\begin{aligned}
\mathsf{E}\left[\mathcal{E}_\phi\left(\hat{f}_{n,\hat{\omega}}, f_\phi^*\right)\right] &\lesssim \inf_{\omega\in\mathcal{A}_n}\left(A_{n,\omega}^{-\alpha/(d-1)} + A_{n,\omega}\frac{\log^3 n}{n}\right) \\
&\le A_{n,\omega_{n,\alpha}}^{-\alpha/(d-1)} + A_{n,\omega_{n,\alpha}}\frac{\log^3 n}{n} \\
&\le \left(\frac{n}{\log^3 n}\right)^{-\frac{d-1}{w_{n,\alpha}+d-1}\frac{\alpha}{d-1}} + \left(\frac{\log^3 n}{n}\right)^{1-\frac{d-1}{w_{n,\alpha}+d-1}} \\
&= \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha}{d-1}\left(1-\frac{k_{n,\alpha}}{\log n}\right)} + \left(\frac{\log^3 n}{n}\right)^{\frac{k_{n,\alpha}}{\log n}}.
\end{aligned}$$

By the definition of $k_{n,\alpha}$, $1 - (k_{n,\alpha}/\log n) \ge (d-1)/(\alpha+d-1)$ and $k_{n,\alpha}/\log n \ge \alpha/(\alpha+d-1) - 1/\log n$. Therefore

$$\mathsf{E}\left[\mathcal{E}_\phi\left(\hat{f}_{n,\hat{\omega}}, f_\phi^*\right)\right] \lesssim \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha}{\alpha+d-1}-\frac{1}{\log n}} \lesssim \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha}{\alpha+d-1}},$$

where the last inequality is due to that $(\log^3 n/n)^{-\frac{1}{\log n}} \lesssim 1$. Since $\mathcal{E}(f, C^*) \le \mathcal{E}_\phi(f, f_\phi^*)$ for any $f \in \mathcal{F}$, we complete the proof. □

1.7.9 Proof of Theorem 1.5.1

Lemma 1.7.13. *Let* $f^*_{\phi,0} = \operatorname{argmin}_{f\in\mathcal{F}_0}\mathcal{E}_\phi(f)$. *There exist positive constants* L_0, N_0, S_0, B_0, F_0 *and* c *such that*

$$\mathcal{E}(f^*_{\phi,0}, C^*) = 0 \tag{1.7.20}$$

and

$$\mathsf{E}\left[\left(\phi(f) - \phi(f^*_{\phi,0})\right)^2\right] \le c\mathcal{E}_\phi(f, f^*_{\phi,0}) \tag{1.7.21}$$

for all $f \in \mathcal{F}_0 = \mathcal{F}^{\mathrm{DNN}}(L_0, N_0, S_0, B_0, F_0)$.

Proof. Let

$$\mathcal{X}^*_{m_0} := \{\mathbf{x} : \operatorname{dist}(\mathbf{x}, D^*) > m_0\}.$$

Using the similar argument used in the proof of Theorem 1.3.4, we can show that there exists a neural network $\tilde{f}^\circ \in \mathcal{F}^{\mathrm{DNN}}(L_0, N_0, S_0, B_0, 1)$ such that $\tilde{f}^\circ(\mathbf{x}) = C^*(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}^*_{m_0}$, for some universal positive constants L_0, N_0, S_0, and B_0. Let F_0 be a positive constant such that $F_0 < \log((1+\eta_0)/(1-\eta_0))$. Define $f^\circ = F_0\tilde{f}^\circ$. Then $f^\circ(\mathbf{x}) = F_0$ whenever $C^*(\mathbf{x}) = 1$ and $f^\circ(\mathbf{x}) = -F_0$ whenever $C^*(\mathbf{x}) = -1$. Since f° has the same sign of the Bayes classifier C^* on $\mathcal{X}^*_{m_0}$, by Assumption M which implies $\mathsf{P}(\mathcal{X}^*_{m_0}) = 1$, we have $\mathcal{E}(f^\circ, C^*) = 0$.

To complete the proof of (1.7.20), we have to show that f° minimizes $\mathcal{E}_\phi(f)$ on $\mathcal{F}_0$. For $\xi \in [0,1]$, let

$$A_\xi(z) := \phi(z)\xi - \phi(-z)(1-\xi).$$

It is easy to see that $A_\xi(z)$ is strictly decreasing on $(-\infty, \log(\xi/(1-\xi)]$ and strictly increasing on $[\log(\xi/(1-\xi)), \infty)$. Hence $F_0 = \operatorname{argmin}_{z\in[-F_0,F_0]} A_\xi(z)$ if $2\xi - 1 > \eta_0$ and $-F_0 = \operatorname{argmin}_{z\in[-F_0,F_0]} A_\xi(z)$ if $2\xi - 1 < -\eta_0$. Note that $f^\circ(\mathbf{x}) = F_0$ if $2\eta(\mathbf{x}) - 1 > \eta_0$ and $f^\circ(\mathbf{x}) = -F_0$ if $2\eta(\mathbf{x}) - 1 < -\eta_0$. By Assumption N, f° minimizes $\mathcal{E}_\phi(f)$ on $\mathcal{F}_0$, and the proof of (1.7.20) is done.

To prove (1.7.21), we fix $\mathbf{x} \in [0,1]^d$ with $\operatorname{dist}(\mathbf{x}, D^*) > m_0$ and $|2\eta(\mathbf{x}) - 1| > \eta_0$ and fix $f \in \mathcal{F}_0^{\mathrm{DNN}}$. Let $z_0 := f^*_{\phi,0}(\mathbf{x})$ and $z := f(\mathbf{x})$. By the Taylor

expansion, we have

$$\phi(yz) - \phi(yz_0) = -\frac{ye^{-yz_0}}{1+e^{-yz_0}}(z-z_0) + \frac{e^{-y\tilde{z}}}{(1+e^{-y\tilde{z}})^2}(z-z_0)^2$$

for any $y \in \{-1,1\}$, where $|\tilde{z}| \in [|z|,|z_0|]$. Let $\bar{\eta} := \eta(\mathbf{x})$ and note that

$$\begin{aligned}\mathsf{E}\left(-\frac{Ye^{-Yz_0}}{1+e^{-Yz_0}}(z-z_0)|\mathbf{X}=\mathbf{x}\right) &= \left[-\bar{\eta}\frac{e^{-z_0}}{1+e^{-z_0}} + (1-\bar{\eta})\frac{e^{z_0}}{1+e^{z_0}}\right](z-z_0)\\ &= \left[\frac{-\bar{\eta}+(1-\bar{\eta})e^{z_0}}{1+e^{z_0}}\right](z-z_0)\end{aligned}$$

Let $\tau := \log((1+\eta_0)/(1-\eta_0)) - F_0 > 0$. If $\bar{\eta} > (1+\eta_0)/2$, $z_0 = F_0$ and hence $z \le z_0$ and

$$\begin{aligned}-\bar{\eta}+(1-\bar{\eta})e^{z_0} &\le -\frac{1+\eta_0}{2} + e^{-\tau}\frac{1-\eta_0}{2}\frac{1+\eta_0}{1-\eta_0}\\ &\le (e^{-\tau}-1)\frac{1+\eta_0}{2} < 0.\end{aligned}$$

On the other hand, if $\bar{\eta} < (1-\eta_0)/2$, $z_0 = -F_0$ and so $z \ge z_0$ and

$$\begin{aligned}-\bar{\eta}+(1-\bar{\eta})e^{z_0} &\ge -\frac{1-\eta_0}{2} + e^{\tau}\frac{1+\eta_0}{2}\frac{1-\eta_0}{1+\eta_0}\\ &\le (e^{\tau}-1)\frac{1-\eta_0}{2} > 0.\end{aligned}$$

Therefore,

$$\left[\frac{-\bar{\eta}+(1-\bar{\eta})e^{z_0}}{1+e^{z_0}}\right](z-z_0) \ge \min\{c_1,c_2\}|z-z_0|,$$

where $c_1 := (1-\mathrm{e}^{-\tau})\frac{1+\eta_0}{2}/(1+\mathrm{e}^{F_0}) > 0$ and $c_2 := (\mathrm{e}^{\tau}-1)\frac{1-\eta_0}{2}/(1+\mathrm{e}^{-F_0}) > 0$. To sum up,

$$\begin{aligned}
&\mathsf{E}\left[\phi\left(Yf(\mathbf{x})\right) - \phi\left(Yf^*_{\phi,0}(\mathbf{x})\right)|\mathbf{X}=\mathbf{x}\right] \\
&\geq \min\{c_1,c_2\}\mathsf{E}\left[\left|f(\mathbf{x})-f^*_{\phi,0}(\mathbf{x})\right||\mathbf{X}=\mathbf{x}\right] + c_3\mathsf{E}\left[\left(f(\mathbf{x})-f^*_{\phi,0}(\mathbf{x})\right)^2|\mathbf{X}=\mathbf{x}\right] \\
&\geq \left(\frac{\min\{c_1,c_2\}}{2F_0}+c_3\right)\mathsf{E}\left[\left(f(\mathbf{x})-f^*_{\phi,0}(\mathbf{x})\right)^2|\mathbf{X}=\mathbf{x}\right],
\end{aligned} \tag{1.7.22}$$

where $c_3 := \mathrm{e}^{-F_0}/(1+\mathrm{e}^{-F_0})^2$. Taking expectation of the both sides with respect to P_X, we obtain the desired result. □

Lemma 1.7.14. *Suppose that $\eta \in \mathcal{H}^{\beta,r}([0,1]^d)$. Let ϕ be the logistic loss. Let $\mathcal{F}_0^{\mathrm{DNN}} = \mathcal{F}^{\mathrm{DNN}}(L_0, N_0, S_0, B_0, F_0)$, where all the constants do not depend on the sample size n. Consider the empirical logistic risk minimizer*

$$\hat{f}^{\mathrm{DNN}}_{\phi,n} = \operatorname*{argmin}_{f\in\mathcal{F}_0^{\mathrm{DNN}}} \frac{1}{n}\sum_{i=1}^{n}\phi(y_i f(\mathbf{x}_i)).$$

Let $f^*_{\phi,0} = \operatorname{argmin}_{f\in\mathcal{F}_0^{\mathrm{DNN}}}\mathcal{E}_\phi(f)$. Let $A>0$ and $\kappa>0$. Then

$$\mathsf{P}\left(\mathcal{E}_\phi(\hat{f}^{\mathrm{DNN}}_{\phi,n}, f^*_{\phi,0}) \geq A\frac{\log^{1+\kappa} n}{n}\right) \lesssim \exp(-C\log^{1+\kappa} n),$$

for some universal constant $C>0$.

Proof. We will check the conditions provided in Section 1.7.2. (A1) and (A3) hold trivially. Since $f^*_{\phi,0} \in \mathcal{F}_0^{\mathrm{DNN}}$, (A2) holds with $a_n = 0$. (A4) is met with $\nu = 1$ by Lemma 1.7.13. Moreover, letting $\delta_n = A\log^{1+\kappa} n/n$, we have that

$$\log\mathcal{N}(\delta_n, \mathcal{F}_0^{\mathrm{DNN}}, \|\cdot\|_\infty) \lesssim \log\left(\frac{n}{\log^{1+\kappa} n}\right) \lesssim \log n \lesssim n\delta_n, \tag{1.7.23}$$

which implies (A5). Therefore following the lines of the proof of Theorem 1.7.3, we obtain

$$\mathsf{P}\left(\mathcal{E}_\phi(\hat{f}_{\phi,n}^{\mathrm{DNN}}, f_{\phi,0}^*) \geq A\frac{\log^{1+\kappa} n}{n}\right) \lesssim \exp(-c\log^{1+\kappa} n),$$

for some constant $c > 0$. □

Proof of Theorem 1.5.1. Since $F_0^{-1} f_{\phi,0}^* = C^*$, P_X-a.s. by Lemma 1.7.13,

$$\begin{aligned}\mathcal{E}(f, f_{\phi,0}^*) = \mathcal{E}(F_0^{-1} f, F_0^{-1} f_{\phi,0}^*) &\leq \mathcal{E}_\phi(F_0^{-1} f, F_0^{-1} f_{\phi,0}^*) \\ &\leq F_0^{-1}\mathsf{E}(|f(\mathbf{X}) - f_{\phi,0}^*(\mathbf{X})|).\end{aligned}$$

In addition, by (1.7.22), we have proved that there exists a constant $c_1 > 0$ such that $\mathsf{E}(|f(\mathbf{X}) - f_{\phi,0}^*(\mathbf{X})|) \leq c_1 \mathcal{E}_\phi(f, f_{\phi,0}^*)$ for any $f \in \mathcal{F}_0^{\mathrm{DNN}}$. Hence we have

$$\mathcal{E}(f, C^*) = \mathcal{E}(f, f_{\phi,0}^*) \leq c_1 F_0^{-1} \mathcal{E}_\phi(f, f_{\phi,0}^*).$$

Since $0 \leq \mathcal{E}(f, C^*) \leq 1$,

$$\begin{aligned}\mathsf{E}\left[\mathcal{E}(\hat{f}_{\phi,n}^{\mathrm{DNN}}, C^*)\right] &\leq \frac{\log^{1+\kappa} n}{n} + \mathsf{P}\left(\mathcal{E}(\hat{f}_{\phi,n}^{\mathrm{DNN}}, C^*) \geq \frac{\log^{1+\kappa} n}{n}\right) \\ &\leq \frac{\log^{1+\kappa} n}{n} + \mathsf{P}\left(\mathcal{E}_\phi(\hat{f}_{\phi,n}^{\mathrm{DNN}}, f_{\phi,0}^*) \geq c_1^{-1} F_0 \frac{\log^{1+\kappa} n}{n}\right) \\ &\leq \frac{\log^{1+\kappa} n}{n} + \exp(-c_2 \log^{1+\kappa} n) \lesssim \frac{\log^{1+\kappa} n}{n}\end{aligned}$$

where the third inequality follows from Lemma 1.7.14. □

1.7.10 Proof of Proposition 1.7.9

We divide the proof into two steps. First we give the proof of approximation of the horizon functions, and then using the result, we prove Proposition 1.7.9.

Lemma 1.7.15 (Approximation of horizon functions). *Let $d \geq 2$, $\alpha > 0$ and $r > 0$. For a horizon function $\Psi_{g,j} = \mathbb{1}(x_j \geq g(\mathbf{x}_{-j}))$ and $\xi > 0$, where $g \in \mathcal{H}^{\alpha,r}([0,1]^{d-1})$, $j \in [d]$, define*

$$\mathcal{X}_{g,j,\xi} = \{\mathbf{x} \in [0,1]^d : x_j - g(\mathbf{x}_{-j}) > \xi\} \cup \{\mathbf{x} \in [0,1]^d : x_j - g(\mathbf{x}_{-j}) < 0\}.$$

Then there exists a neural network

$$f^\circ \in \mathcal{F}^{\text{DNN}}\left(L_0 \log(1/\xi), N_0 \xi^{-(d-1)/\alpha}, S_0 \xi^{-(d-1)/\alpha} \log(1/\xi), B_0 \xi^{-1}, 1\right), \tag{1.7.24}$$

where the positive constants L_0, N_0, S_0 and B_0 depend only on d, α, and r, such that

$$\sup_{\mathbf{x} \in \mathcal{X}_{g,j,\xi}} \left| f^\circ(\mathbf{x}) - \Psi_{g,j}(\mathbf{x}) \right| = 0.$$

Proof. Let $\xi > 0$ be given. By Proposition 1.7.6, there is a neural network g° on $[0,1]^{d-1}$ such that $\|g^\circ - g\|_\infty < \xi/4$ with the number of layers $\lesssim \log(1/\xi)$, the maximum number of hidden nodes $\lesssim \xi^{-(d-1)/\alpha}$, sparsity $\lesssim \xi^{-(d-1)/\alpha} \log(1/\xi)$, the largest absolute value ≤ 1, and $\|g^\circ\|_\infty \leq r + 1$. We construct the neural network f°, which approximates $\Psi_{g,j}$, as

$$\begin{aligned} f^\circ(\mathbf{x}) &:= 2\left\{\sigma\left(\frac{2}{\xi}\left(x_j - g^\circ(\mathbf{x}_{-j}) - \frac{\xi}{4}\right)\right) - \sigma\left(\frac{2}{\xi}\left(x_j - g^\circ(\mathbf{x}_{-j}) - \frac{3\xi}{4}\right)\right)\right\} - 1 \\ &= \begin{cases} 1 & \text{if } x_j \geq g^\circ(\mathbf{x}_{-j}) + 3\xi/4 \\ 4(x_1 - g^\circ(\mathbf{x}_{-j}) - \xi/4)/\xi - 1 & \text{if } g^\circ(\mathbf{x}_{-j}) + \xi/4 \leq x_j < g^\circ(\mathbf{x}_{-j}) + 3\xi/4 \\ -1 & \text{if } x_j < g^\circ(\mathbf{x}_{-j}) + \xi/4. \end{cases} \end{aligned}$$

Given $\mathbf{x} \in \mathcal{X}_{g,j,\xi}$, we have:

- when $x_j - g(\mathbf{x}_{-j}) > \xi$,

$$x_j > g^\circ(\mathbf{x}_{-j}) + (g(\mathbf{x}_{-j}) - g^\circ(\mathbf{x}_{-j})) + \xi \geq g^\circ(\mathbf{x}_{-j}) + 3\xi/4,$$

- when $x_j - g(\mathbf{x}_{-j}) < 0$,

$$x_j < g^\circ(\mathbf{x}_{-j}) + (g(\mathbf{x}_{-j}) - g^\circ(\mathbf{x}_{-j})) \leq g^\circ(\mathbf{x}_{-j}) + \xi/4.$$

Hence, $\Psi_{g,j} = f^\circ$ on $\mathcal{X}_{g,j,\xi}$. □

Proof of Proposition 1.7.9. Let $f_{t,k}$ be a neural network in (1.7.24) such that

$$\sup_{\mathbf{x}\in\mathcal{X}_{g_{t,k},j_{t,k},\xi}} |f_{t,k}(\mathbf{x}) - \Psi_{g_{t,k},j_{t,k}}| = 0,$$

where

$$\begin{aligned}\mathcal{X}_{g_{t,k},j_{t,k},\xi} :=& \{\mathbf{x}\in[0,1]^d : x_{j_{t,k}} - g_{t,k}(\mathbf{x}_{-j_{t,k}}) > \xi\} \\ &\cup \{\mathbf{x}\in[0,1]^d : x_{j_{t,k}} - g_{t,k}(\mathbf{x}_{-j_{t,k}}) < 0\}.\end{aligned}$$

Define the neural network f° as

$$f^\circ(\mathbf{x}) := \sum_{t=1}^{T} \sigma\left(\sum_{k=1}^{K} f_{t,k}(\mathbf{x}) - (K-2)\right) - 1.$$

Because $(A_t)_{t\in[T]}$ are disjoint,

$$\mathcal{X}_{A_{1:T},\xi} := \bigcap_{t=1}^{T}\left(A_t^c \cup A_{t,\xi}\right) = \left(\bigcap_{t=1}^{T} A_t^c\right) \bigcup \left(\bigcup_{t=1}^{T} A_{t,\xi}\right).$$

where $(A_{t,\xi})_{t\in[T]}$ are defined in (1.7.15). Given $\mathbf{x}\in\mathcal{X}_{A_{1:T},\xi}$, we have the followings.

- Suppose that $\mathbf{x}\in\bigcap_{t=1}^{T} A_t^c$. Then there is $k_t\in[K]$ such that $f_{t,k_t}(\mathbf{x}) = -1$ for each $t\in[T]$. Hence, $\sum_{k=1}^{K} f_{t,k}(\mathbf{x}) \le K-2$ for every $t\in[T]$ and thus $f^\circ(\mathbf{x}) = -1$. Also it is obvious that $C(\mathbf{x}) = -1$.
- Suppose that $\mathbf{x}\in A_{t,\xi}$. Then $f_{t,k}(\mathbf{x}) = 1$ for any $k\in[K]$, which implies $\sigma\left(\sum_{k=1}^{K} f_{t,k}(\mathbf{x}) - (K-2)\right) = 2$. On the other hand, since $A_{t,\xi}\subset A_t\subset A_{t'}^c$ for every $t'\neq t$, it follows that $\sum_{k=1}^{K} f_{t',k}(\mathbf{x}) \le K-2$ and so $\sigma\left(\sum_{k=1}^{K} f_{t,k}(\mathbf{x}) - (K-2)\right) = 0$ for every $t'\neq t$. Thus $f^\circ(\mathbf{x}) = 1$. Also we have that $C(\mathbf{x}) = 1$ since $A_{t,\xi}\subset A_t$.

Thus, $f^\circ = C$ on $\mathcal{X}_{A_{1:T},\xi}$. □

Chapter 2

Rate-optimal sparse learning for deep neural networks

2.1 Introduction

Sparse learning of deep neural networks (DNN) has received much attention in artificial intelligence and statistics. In artificial intelligence, there are a lot of evidence [43, 34, 62] to support that sparse DNN can reduce the complexity of a leaned DNN significantly (in terms of the number of parameters as well as the numbers of hidden layers and hidden nodes) without hampering prediction accuracy much. By doing so, we can reduce memory and energy consumption at the prediction phase.

In statistics, recent studies about DNNs for nonparametric regression and classification [80, 46, 87, 7, 51] have proved that a DNN estimator minimizing an empirical risk with a certain sparsity constraint achieves the minimax optimality for a wide class of functions including smooth functions, piecewise smooth functions and smooth decision boundaries. However, there are still two unanswered problems. The first problem is to choose a suitable level of sparsity, which depends on the unknown smoothness and/or the unknown intrinsic dimensionality of the true function. The second one is computation. Learning a deep architecture with a given sparsity constraint is computationally intractable since we need to explore a large number of possible configurations of sparsity pattern in the network parameter.

In this chapter, we propose a novel learning method of sparse DNNs for

nonparametric regression and classification, which answers the two problems mentioned above in the sparsity-constrained empirical risk minimization (ERM) algorithm. The proposed learning algorithm is to learn a DNN by minimizing the penalized empirical risk, which is the sum of the empirical risk and the clipped L_1 penalty [104]. By choosing the position of the clipping carefully, we establish an oracle inequality for the excess risk of the proposed sparse DNN estimator and derive convergence rates for several learning tasks. In particular, the proposed sparse DNN estimator can adaptively attain minimax convergence rates for various nonparametric regression problems.

Although nonconvex penalties such as the clipped L_1 penalty are popular for high-dimensional linear regressions [29, 102], they are seldom used for DNN. Instead, L_1 norm-based penalties such as Lasso and Group Lasso are popular [61, 97]. This would be partly because of the convexity of the L_1 penalty. For computation with the clipped L_1 penalty, we develop an optimization algorithm adopting the concave-convex procedure (CCCP) [101], which has been popularly used for nonconvex penalized linear regression problems [e.g., 50].

2.1.1 Notation

We denote by $\mathbb{1}(\cdot)$ the indicator function. Let $\mathbb{R}$ be the set of real numbers and $\mathbb{N}$ be the set of natural numbers. Let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Let $[m] := \{1, 2, \ldots, m\}$ for $m \in \mathbb{N}$. For two real numbers a and b, we write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. For a real valued vector $\mathbf{x} \equiv (x_1, \ldots, x_d) \in \mathbb{R}^d$, we let $\|\mathbf{x}\|_0 := \sum_{j=1}^d \mathbb{1}(x_j \neq 0)$, $\|\mathbf{x}\|_p := \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$ for $p \in [1, \infty)$ and $\|\mathbf{x}\|_\infty := \max_{1 \leq j \leq d} |x_j|$. For a real-valued function $f : \mathcal{X} \mapsto \mathbb{R}$, we let $\|f\|_{\infty, \mathcal{X}} := \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})|$. If the domain of the function is clear in the context, we omit the subscript $\mathcal{X}$ to write $\|f\|_\infty := \|f\|_{\infty, \mathcal{X}}$. For $p \in [1, \infty)$ and a distribution Q on $\mathcal{X}$, let $\|f\|_{p, \mathsf{Q}} := \left(\int |f(\mathbf{x})|^p \mathrm{d}\mathsf{Q}(\mathbf{x})\right)^{1/p}$.

2.1.2 Deep neural networks

A DNN with $L \in \mathbb{N}$ layers, $N_l \in \mathbb{N}$ many nodes at the l-th hidden layer for $l = 1, \dots, L$, input of dimension N_0, output of dimension N_{L+1} and nonlinear activation function $\rho : \mathbb{R} \mapsto \mathbb{R}$ is expressed as

$$f(\mathbf{x}) = A_{L+1} \circ \rho_L \circ A_L \circ \cdots \circ \rho_1 \circ A_1(\mathbf{x}), \tag{2.1.1}$$

where $A_l : \mathbb{R}^{N_{l-1}} \mapsto \mathbb{R}^{N_l}$ is an affine linear map defined by $A_l(\mathbf{x}) = \mathbf{W}_l \mathbf{x} + \mathbf{b}_l$ for given $N_l \times N_{l-1}$ dimensional weight matrix $\mathbf{W}_l$ and N_l dimensional bias vector $\mathbf{b}_l$, and $\rho_l : \mathbb{R}^{N_l} \mapsto \mathbb{R}^{N_{l+1}}$ is an element-wise nonlinear activation map defined as $\rho_l(\mathbf{z}) := (\rho(z_1), \dots, \rho(z_{N_l}))^\top$. We let $\boldsymbol{\theta}(f)$ denote a parameter, which is a concatenation of all the weight matrices and the bias vectors, of the DNN f, that is,

$$\boldsymbol{\theta}(f) := (\mathrm{vec}(\mathbf{W}_1)^\top, \mathbf{b}_1^\top, \dots, \mathrm{vec}(\mathbf{W}_{L+1})^\top, \mathbf{b}_{L+1}^\top)^\top,$$

where $\mathrm{vec}(\mathbf{W})$ transforms the matrix $\mathbf{W}$ into the corresponding vector by concatenating the column vectors. We call $\boldsymbol{\theta}(f)$ the parameter of the DNN f.

We let $\mathcal{F}^{\mathrm{DNN}}_{\rho,d,o}$ be the class of DNNs which take d-dimensional input (i.e. $N_0 = d$) to produce o-dimensional output (i.e. $N_{L+1} = o$) and use the activation function $\rho : \mathbb{R} \mapsto \mathbb{R}$. In this chapter, we focus on the real-valued DNNs, i.e., $o = 1$, but the results in this chapter can be extended easily for the case of $o \geq 2$.

For a given DNN f, we let $\mathsf{depth}(f)$ denote the depth (i.e., the number of hidden layers) and $\mathsf{width}(f)$ denote the width (i.e., the maximum of the numbers of hidden nodes at each layer) of the DNN f. Throughout this chapter, we consider a class of DNNs with some constraints on the architecture, parameter and output value of a DNN such that

$$\begin{aligned}\mathcal{F}^{\mathrm{DNN}}_{\rho}(L, N, B, F) := \Big\{ f \in \mathcal{F}^{\mathrm{DNN}}_{\rho,d,1} :& \mathsf{depth}(f) \leq L, \mathsf{width}(f) \leq N - 1, \\ & \|\boldsymbol{\theta}(f)\|_\infty \leq B, \|f\|_\infty \leq F \\ & \rho \text{ is 1-Lipschitz} \Big\}.\end{aligned} \tag{2.1.2}$$

for positive constants L, N, B and F. Here we omit the input dimension to simplify the notation. The Lipschitzness assumption on the activation function ρ is needed to bound the covering number of the class of DNNs. The ReLU activation function $x \mapsto \max\{0, x\}$ and the sigmoid activation function $x \mapsto 1/(1 + \mathrm{e}^{-x})$, which are the two most widely used activation functions, are both Lipschitz.

2.1.3 Empirical risk minimization algorithm with a sparsity constraint and its nonadaptiveness

Most studies about DNNs for nonparametric regression [80, 46, 87, 7] considered the ERM algorithm with a certain sparsity constraint, which is summarized as follows. Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be n many input-output pairs which are assumed to be independent and identical random vectors distributed according to P on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. First, a class of sparsity constrained DNNs with the sparsity level $S \in \mathsf{N}$ is defined as

$$\mathcal{F}_\rho^{\mathrm{DNN}}(L, N, B, F, S) := \Big\{ f \in \mathcal{F}_\rho^{\mathrm{DNN}}(L, N, B, F) : \|\boldsymbol{\theta}(f)\|_0 \le S \Big\}. \tag{2.1.3}$$

Then for a given loss $l : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$, the sparsity-constrained ERM estimator is given by

$$\hat{f}_n^{\mathrm{ERM}} \in \operatorname*{argmin}_{f \in \mathcal{F}_\rho^{\mathrm{DNN}}(L_n, N_n, B_n, F_n, S_n)} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(\mathbf{X}_i)). \tag{2.1.4}$$

with suitably chosen architecture parameters L_n, N_n, B_n, F_n and sparsity S_n.

Many studies [80, 46, 87, 7] have been proved that the estimator $\hat{f}_n^{\mathrm{ERM}}$ can attain minimax optimality in various supervised learning tasks, but most results are nonadaptive. To be more specific, let $f_\ell^\star := \operatorname{argmin}_{f \in \mathcal{F}} \mathsf{E}\ell(Y, f(\mathbf{X}))$, where $\mathcal{F}$ is a set of all real-valued measurable function on $\mathcal{X}$. Define the excess risk of a function f as

$$\mathcal{E}_{\mathsf{P}}(f) := \mathsf{E}\ell(Y, f(\mathbf{X})) - \mathsf{E}\ell(Y, f_\ell^\star(\mathbf{X})).$$

If ℓ is the square loss, the activation function is the ReLU and $f_\ell^\star$ belongs to the class of Hölder functions of smoothness $\alpha > 0$ with radius R (see (2.3.8) in 2.3 for the definition of Hölder functions), Schmidt-Hieber [80] proves that the convergence rate of the excess risk $\mathcal{E}_{\mathsf{P}}(\hat{f}_n^{\mathrm{ERM}})$ is $O(n^{-\frac{2\alpha}{2\alpha+d}} \log^3 n)$, which is is minimax optimal upto a logarithm factor, provided that $L_n \lesssim \log n$, $N_n \lesssim n^{c_1}$, $B_n \lesssim n^{c_2}$ and $S_n \asymp n^{\frac{d}{2\alpha+d}} \log n$ for some positive constants c_1 and c_2. That is, the sparsity level S_n for attaining the minimax optimality depends on the smoothness of the true function $f_\ell^\star$ which is unknown. This nonadaptiveness is also unavoidable for classification. For details, see [51].

2.1.4 Outline

This chapter is organized as follows. In Section 2.2, we propose a sparse learning method for DNNs. In Section 2.3, we provide the oracle inequalities for the proposed sparse DNN estimator. Based on these oracle inequalities, we derive convergence rates of our estimator for several supervised learning problems. In Section 2.4, we present an optimization algorithm for our estimator. In Section 2.5, we conduct numerical study to assess the finite-sample performance of our estimator. Concluding remarks follow in Section 2.6, and the proofs are gathered in Section 2.7.

2.2 Learning sparse deep neural networks with the clipped L_1 penalty

In this chapter, we consider the penalized empirical risk minimizer over DNNs, which is obtained by

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}_n^{\mathrm{DNN}}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell\left(Y_i, f(\mathbf{X}_i)\right) + J_n(f) \right], \tag{2.2.1}$$

where $\mathcal{F}_n^{\mathrm{DNN}}$ is a certain class of DNNs and $J_n(f)$ is a sparsity-inducing penalty function. We call $\hat{f}_n$ the sparse-penalized DNN estimator. For the sparsity-inducing penalty $J_n(f)$, we propose to use the clipped L_1 penalty

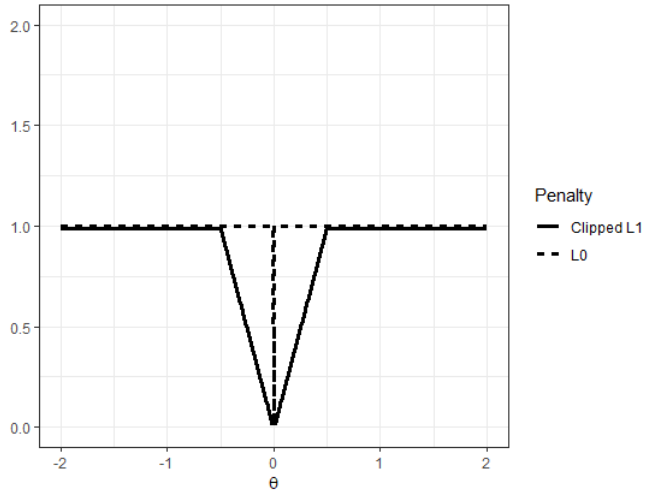


FIGURE 2.1: The clipped L_1 and L_0 penalties

given by

$$J_n(f) := J_{\lambda_n,\tau_n}(f) := \lambda_n \big\|\boldsymbol{\theta}(f)\big\|_{\mathsf{Clip},\tau_n}, \tag{2.2.2}$$

for tuning parameters $\lambda_n > 0$ and $\tau_n > 0$, where $\|\cdot\|_{\mathsf{Clip},\tau}$ denotes the *clipped* L_1 *norm* with a clipping threshold $\tau > 0$ [104] defined as

$$\|\boldsymbol{\theta}\|_{\mathsf{Clip},\tau} := \sum_{j=1}^{p} \left(\frac{|\theta_j|}{\tau} \wedge 1 \right) \tag{2.2.3}$$

for a p-dimensional vector $\boldsymbol{\theta} \equiv (\theta_j)_{j\in[p]}$.

The clipped L_1 norm can be viewed as a continuous relaxation of the L_0 norm $\|\boldsymbol{\theta}\|_0$. Figure 2.1 shows the L_0 and the clipped L_1 norms. The continuity of the clipped L_1 norm makes the optimization (2.2.1) much easier than the optimization of the L_0 norm, which will be discussed in Section 2.4.

The main result of this chapter is that with suitable choices for λ_n and τ_n, which do depend on neither training data nor the true distribution, the sparse-penalized DNN estimator (2.2.1) with the clipped L_1 penalty can adaptively attain minimax optimality without knowing (nonadaptive) sparsity constraint.

2.3 Main results

In this section, we provide theoretical justification of our sparse-penalized DNN estimator (2.2.1) in both regression and binary classification tasks. We prove that minimax optimal converge rates of the excess risk can be obtained adaptively for various nonparametric regression and classification tasks.

2.3.1 Nonparametric regression

We first consider a nonparametric regression task, where the response $Y \in \mathbb{R}$ and input $\mathbf{X} \in [0,1]^d$ is generated from the model

$$Y = f^\star(\mathbf{X}) + \epsilon, \quad \mathbf{X} \sim \mathsf{P}_{\mathbf{X}} \tag{2.3.1}$$

where $f^\star : [0,1]^d \mapsto \mathbb{R}$ is the unknown true regression function, $\mathsf{P}_{\mathbf{X}}$ is a distribution on $[0,1]^d$ and ϵ is an error variable independent to the input variable $\mathbf{X}$. For technical simplicity, we focus on the sub-Gaussian error such that

$$\mathsf{E}(\mathrm{e}^{a\epsilon^2}) < \infty \tag{2.3.2}$$

for some $a > 0$. We denote by $\mathcal{P}_{a,F^\star}$ the set of distributions of $(\mathbf{X}, Y)$ satisfying the model (2.3.1):

$$\mathcal{P}_{a,F^\star} := \left\{ \text{Model (2.3.1)} : \mathsf{E}(\mathrm{e}^{a\epsilon^2}) < \infty, \|f^\star\|_\infty \le F^\star \right\}.$$

The problem is to estimate the unknown true regression function $f^\star$ from given training data $((\mathbf{X}_1, Y_i))_{i\in[n]} \sim \mathsf{P}^n$. We evaluate the performance of an estimator $\hat{f}$ by the expected $L_2(\mathsf{P}_{\mathbf{X}})$ error

$$\mathsf{E}\left[\|\hat{f} - f^\star\|_{2,\mathsf{P}_{\mathbf{X}}}^2 \right],$$

where the expectation is taken over the training data and

$$\|\hat{f} - f^\star\|_{2,\mathsf{P}_{\mathbf{X}}}^2 := \int |\hat{f}(\mathbf{x}) - f^\star(\mathbf{x})|^2 \mathrm{d}\mathsf{P}_{\mathbf{X}}(\mathbf{x}).$$

The following theorem provides an oracle inequality for the expected $L_2(\mathsf{P}_{\mathbf{X}})$ error of the sparse-penalized DNN estimator.

Theorem 2.3.1. *Assume that the true generative model P is in $\mathcal{P}_{a,F^\star}$. Let $\mathcal{F}_n^{\mathrm{DNN}} := \mathcal{F}_\rho^{\mathrm{DNN}}(L_n, N_n, B_n, F)$ and assume that $1 \le L_n \lesssim \log n$, $2 \le N_n \lesssim n$, $1 \le B_n \lesssim n^b$ for some $b > 0$ and $F \ge F^\star$. Then the sparse-penalized DNN estimator*

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}_n^{\mathrm{DNN}}} \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i))^2 + J_{\lambda_n, \tau_n}(f) \right], \tag{2.3.3}$$

where

$$J_{\lambda_n, \tau_n}(f) := \lambda_n \big\| \boldsymbol{\theta}(f) \big\|_{\mathsf{Clip}, \tau_n}, \tag{2.3.4}$$

with $\lambda_n \gtrsim \log^5 n / n$ and $\tau_n \lesssim n^{-1} L_n^{-1} (N_n B_n)^{-L_n}$, satisfies

$$\mathsf{E}\left[\|\hat{f}_n - f^\star\|_{2,\mathsf{P}_{\mathbf{X}}}^2 \right] \le 2 \inf_{f \in \mathcal{F}_n^{\mathrm{DNN}}} \left\{ \|f - f^\star\|_{2,\mathsf{P}_{\mathbf{X}}}^2 + J_{\lambda_n, \tau_n}(f) \right\} + c \frac{\log^2 n}{n} \tag{2.3.5}$$

for some $c > 0$, where the expectation is taken over the training data.

Proof. The proof is deferred to Section 2.7.2. □

The following theorem, which is a direct consequence of Theorem 2.3.1, provides a useful tool to derive convergence rates of the sparse-penalized DNN estimator for various classes of functions to which the true regression function $f^\star$ belongs.

Theorem 2.3.2. *Let the class of DNNs $\mathcal{F}_n^{\mathrm{DNN}} := \mathcal{F}_\rho^{\mathrm{DNN}}(L_n, N_n, B_n, F)$ and the penalty function $J_{\lambda_n, \tau_n}(\cdot)$ be as in Theorem 2.3.1. Let $\mathcal{F}^\star$ be a set of some real-valued functions on $[0,1]^d$. Define*

$$\mathcal{F}_{0,n}^{\mathrm{DNN}}(S) := \left\{ f \in \mathcal{F}_n^{\mathrm{DNN}} : \|\boldsymbol{\theta}(f)\|_0 \le S \right\}$$

for $S > 0$. Assume that there are universal constants $\kappa > 0$, $r > 0$ and $c > 0$ such that

$$\sup_{f^\diamond \in \mathcal{F}^\star} \inf_{f \in \mathcal{F}_{0,n}^{\mathrm{DNN}}(S)} \|f - f^\diamond\|_{2,\mathsf{P}_{\mathbf{X}}} \le c \left(\frac{S}{\log^r n} \right)^{-\kappa}. \tag{2.3.6}$$

for any $S > 0$ and $n \in \mathbb{N}$. Then the sparse-penalized DNN estimator defined by (2.3.3) satisfies

$$\sup_{\mathsf{P} \in \mathcal{P}_{a,F^\star} : f^\star \in \mathcal{F}^\star} \mathsf{E}\left[\|\hat{f}_n - f^\star\|_{2,\mathsf{P}_X}^2 \right] \lesssim n^{-\frac{2\kappa}{2\kappa+1}} \log^{5+r} n. \tag{2.3.7}$$

Proof. The proof is deferred to Section 2.7.3. □

The choice of hyperparemters λ_n and τ_n depend only on sample size n. That is, the proposed sparse DNN estimator can automatically choose the suitable sparsity without the knowledge of the true regression function and thus can attain the optimal convergence adaptively. Below, we list up the examples where the sparse penalized DNN can attain the minimax optimal convergence rate (upto a logarithm factor) adaptively.

Hölder functions The Hölder space of smoothness $\alpha > 0$ with radius $R > 0$ is defined as

$$\mathcal{H}^{\alpha,R}(\mathcal{X}) := \left\{ f : \mathcal{X} \mapsto \mathbb{R} : \|f\|_{\mathcal{H}^\alpha(\mathcal{X})} \leq R \right\}, \tag{2.3.8}$$

where $\|f\|_{\mathcal{H}^\alpha(\mathcal{X})}$ denotes the Hölder norm defined by

$$\begin{aligned} \|f\|_{\mathcal{H}^\alpha(\mathcal{X})} := & \sum_{\mathbf{m} \in \mathbb{N}_0^d : \|\mathbf{m}\|_1 \leq \lfloor \alpha \rfloor} \|\partial^{\mathbf{m}} f\|_\infty \\ & + \sum_{\mathbf{m} \in \mathbb{N}_0^d : \|\mathbf{m}\|_1 = \lfloor \alpha \rfloor} \sup_{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, \mathbf{x}_1 \neq \mathbf{x}_2} \frac{|\partial^{\mathbf{m}} f(\mathbf{x}_1) - \partial^{\mathbf{m}} f(\mathbf{x}_2)|}{|\mathbf{x}_1 - \mathbf{x}_2|^{\alpha - \lfloor \alpha \rfloor}} \end{aligned}$$

where $\partial^{\mathbf{m}} f$ denotes the partial derivative of f of order $\mathbf{m}$.

Yarotsky [100] and Schmidt-Hieber [80] proved that for $\mathcal{H}^{\alpha,R}([0,1]^d)$, the class of DNNs $\mathcal{F}_\rho^{\mathrm{DNN}}(L_n, N_n, 1, F)$ with $L_n \asymp \log n$, $N_n \asymp n^{\frac{d}{2\alpha+d}}$, $F > F^\star$ and ρ being the ReLU satisfies the condition (2.3.6) with $\kappa = \alpha/d$ and $r = 1$. Hence, Theorem 2.3.2 implies that the convergence rate of the sparse-penalized DNN estimator defined by (2.3.3) is

$$n^{-\frac{2\alpha}{2\alpha+d}} \log^6 n, \tag{2.3.9}$$

which is the minimax optimal up to a logarithm factor.

Also, Ohn and Kim [71] (see Theorem A.4.1 in Appendix A) proved that the same result holds for the quite general family of activation functions, which includes most of the widely used activation functions such as ReLU, LeakyReLU, sigmoid and tanh. Hence if the true regression function is in $\mathcal{H}^{\alpha,R}([0,1]^d)$, the sparse-penalized DNN estimator defined by (2.3.3) with the general activation function considered in [71] (for details, see Section A.3) attains the same rate $n^{-\frac{2\alpha}{2\alpha+d}}\log^6 n$. The curse of dimensionality in the above rate can be relaxed by some structural assumption on the regression function, which is considered in the next example.

Composition structured functions Schmidt-Hieber [80] considered so-called composition structured regression functions which include generalized additive models and sparse tensor decomposition models. This class is defined as

$$\Big\{f = g_q \circ \cdots \circ g_1 : g_j = (g_{j,k})_{k\in[d_{j+1}]} : [a_j,b_j]^{d_j} \mapsto [a_{j+1},b_{j+1}]^{d_{j+1}}, \\ g_{j,k} \in \mathcal{H}^{\alpha_j,R}([a_j,b_j]^{t_j})\Big\} \tag{2.3.10}$$

where $d_j \in \mathbb{N}$, $t_j \in [d_j]$ and $\alpha_j > 0$ for $j \in [q]$ with $d_1 = d$, $d_{q+1} = 1$, and $\max_{j\in[q+1]} |a_j| \vee |b_j| \le K$ for some $K > 0$.

Schmidt-Hieber [80] provided a DNN approximation result of a composition structured function. This result is described by the effective smoothness defined by

$$\alpha_j^* = \alpha_j \prod_{h=j+1}^{q} (\alpha_h \wedge 1).$$

Schmidt-Hieber [80] showed that for the class of composition structured functions (2.3.10), the class of DNNs $\mathcal{F}_\rho^{\mathrm{DNN}}(L_n, N_n, 1, F)$ with

$$L_n \asymp \log n, \quad N_n \asymp \max_{j=0,1,\dots,q} n^{\frac{t_j}{2\alpha_j^*+t_j}}, \quad F > F^\star$$

and ρ being the ReLU satisfies the condition (2.3.6) with

$$\kappa = \min_{j=0,1,\ldots,q} \frac{\alpha_j^*}{t_j},$$

and $r = 1$. Thus Theorem 2.3.2 implies that the sparse-penalized DNN estimator defined by (2.3.3) with the ReLU activation function attains the rate

$$\max_{j=0,1,\ldots,q} n^{-\frac{2\alpha_j^*}{2\alpha_j^*+t_j}} \log^6 n, \tag{2.3.11}$$

which is minimax optimal up to a logarithmic factor.

Piecewise smooth functions Petersen and Voigtlaender [76] and Imaizumi and Fukumizu [46] introduced a notion of piecewise smooth functions, which has support divided into several pieces with smooth boundaries and smooth only within each of the pieces. Formally, the class of the piecewise smooth functions is defined as

$$\begin{aligned}\Big\{ f = \sum_{m=1}^{M} g_m \prod_{k\in[K]} \mathbb{1}(x_{j_{m,k}} \geq h_{m,k}(\mathbf{x}_{-j_{m,k}})) : \\ g_m \in \mathcal{H}^{\alpha,R}([0,1]^d), h_{m,k} \in \mathcal{H}^{\beta,R}([0,1]^{d-1}) \Big\}\end{aligned} \tag{2.3.12}$$

for some $M \in \mathbb{N}$, $K \in \mathbb{N}$, $\alpha > 0$, $\beta > 0$, $R > 0$ and $j_{m,k} \in [d]$ for $m \in [M]$ and $k \in [K]$.

Petersen and Voigtlaender [76] and Imaizumi and Fukumizu [46] showed that for the class of piecewise smooth functions (2.3.12), the class of DNNs $\mathcal{F}_\rho^{\mathrm{DNN}}(L_n, N_n, B_n, F)$ with

$$L_n \asymp \log n, \quad N_n \asymp n^{\frac{d}{2\alpha+d}} \vee n^{\frac{d-1}{\beta+d-1}}, \quad B_n \asymp n, \quad F > F^\star$$

and ρ being the ReLU satisfies the condition (2.3.6) with

$$\kappa = \frac{\alpha}{d} \wedge \frac{\beta}{2(d-1)}$$

and $r = 1$, provided that the marginal distribution $\mathsf{P}_{\mathbf{X}}$ of the input variable admits a density $\frac{\mathrm{d}\mathsf{P}_{\mathbf{X}}}{\mathrm{d}\mu}$ with respect to the Lebesgue measure μ such that $\sup_{\mathbf{x}\in[0,1]^d} \frac{\mathrm{d}\mathsf{P}_{\mathbf{X}}}{\mathrm{d}\mu}(\mathbf{x}) \leq p_0$ for some $p_0 > 0$. Hence Theorem 2.3.2 implies that the sparse-penalized DNN estimator defined by (2.3.3) with the ReLU activation function attains the rate

$$\left\{ n^{-\frac{2\alpha}{2\alpha+d}} \vee n^{-\frac{\beta}{\beta+d-1}} \right\} \log^6 n, \tag{2.3.13}$$

which is minimax optimal up to a logarithmic factor.

Besov and mixed smooth Besov functions Suzuki [87] proved that the DNN estimator minimizing an empirical risk with a certain sparsity constraint is minimax optimal for the estimation of regression function in a Besov space, which is a certain generalization of the Hölder space. The minimax optimal rate of estimation over the Besov space with smoothness α is given by

$$n^{-\frac{2\alpha}{2\alpha+d}}$$

and it can be shown that the sparse-penalized DNN estimator defined by (2.3.3) with the ReLU activation function attains the above rate as done in the previous examples.

The dependency on the dimension d of the above rate can be avoided by introducing the mixed smoothness. The mixed smooth Besov space is a function space defined via the notion of mixed smoothness which measures how the function is smooth in a coordinate-wise manner. For the detailed definition, we refer to [87]. Theorem 1 of [87] showed that for the mixed smooth Besov space with mixed smoothness α, the class of DNNs $\mathcal{F}_{\rho}^{\mathrm{DNN}}(L_n, N_n, B_n, F)$ with

$$L_n \asymp \log n, \quad N_n \asymp n^{\frac{1}{2\alpha+1}}, \quad B_n \asymp n, \quad F > F^{\star}$$

and ρ being the ReLU satisfies the condition (2.3.6) with $\kappa = \alpha$ and $r = d$. Hence, Theorem 2.3.2 implies that the sparse-penalized DNN estimator

defined by (2.3.3) with the ReLU activation function attains the rate

$$n^{-\frac{2\alpha}{2\alpha+1}} \log^{5+d} n \tag{2.3.14}$$

which is minimax optimal up to a logarithmic factor.

2.3.2 Classification with strictly convex losses

In this section, we consider a binary classification problem. The goal of classification is to find a real-valued function f (called a classification function) such that $f(\mathbf{X})$ is a good prediction of the label $Y \in \{-1, 1\}$ for a new sample $(\mathbf{X}, Y)$. In practice, the margin-based loss function, which evaluates the quality of the prediction by f for a sample $(\mathbf{X}, Y)$ based on its margin $Yf(\mathbf{X})$, is popularly used. Examples of the margin based loss functions are the 0-1 loss $\mathbb{1}(Yf(\mathbf{X}) < 0)$, hinge loss $(1 - (Yf(\mathbf{X})) \vee 0$, exponential loss $\exp(-(Yf(\mathbf{X}))$ and logistic loss $\log(1 + \exp(-(Yf(\mathbf{X}))))$. Here we focus on strictly convex losses including the exponential and the logistic losses. In particular, the logistic loss is most widely used for learning a DNN classifier in practice.

We assume that the label $Y \in \{-1, 1\}$ and input $\mathbf{X} \in [0, 1]^d$ is generated from the model

$$Y|\mathbf{X} = \mathbf{x} \sim 2\mathsf{Bernoulli}(\eta(\mathbf{x})) - 1, \quad \mathbf{X} \sim \mathsf{P}_{\mathbf{X}} \tag{2.3.15}$$

where $\eta(\mathbf{x})$ is called a conditional class function and $\mathsf{P}_{\mathbf{X}}$ is a distribution on $[0, 1]^d$. Our aim is to find a real-valued function f so that the *excess risk* of f given below as close to zero as possible:

$$\mathcal{E}_{\mathsf{P}}(f) := \mathsf{E}(\ell(Yf(\mathbf{X}))) - \mathsf{E}(\ell(Yf_\ell^\star(\mathbf{X}))),$$

where ℓ is a margin-based loss function, $f_\ell^\star = \operatorname{argmin}_{f \in \mathcal{F}} \mathsf{E}(\ell(Yf(\mathbf{X})))$ is the optimal classification function and $\mathcal{F}$ is a set of all real-valued measurable functions on $[0, 1]^d$. We assume that $\|f_\ell^\star\|_\infty \leq F^\star$ for some $F^\star > 0$. This assumption is satisfied if the conditional class probability $\eta(\mathbf{x})$ satisfies $\inf_{\mathbf{x} \in [0,1]^d} \eta(\mathbf{x}) \wedge (1 - \eta(\mathbf{x})) \geq \eta_0$ for some $\eta_0 > 0$, i.e., η is bounded away from 0 and 1, for the exponential or logistic loss. We denote by $\mathcal{Q}_{F^\star}$ the set

of distributions satisfying the above assumption, that is,

$$\mathcal{Q}_{F^\star} = \left\{\text{Model (2.3.15)} : \|f_\ell^\star\|_\infty \le F^\star\right\}.$$

The following theorem states the oracle inequality for the excess risk of a DNN estimated with the clipped L_1 penalty.

Theorem 2.3.3. *Let ℓ be a strictly convex margin-based loss function with continuous first and second derivatives. Assume the true generative model P is in $\mathcal{Q}_{F^\star}$. Let $\mathcal{F}_n^{\mathrm{DNN}} := \mathcal{F}_\rho^{\mathrm{DNN}}(L_n, N_n, B_n, F)$ and assume that $1 \le L_n \lesssim \log n$, $2 \le N_n \lesssim n$, $1 \le B_n \lesssim n^b$ for some $b > 0$ and $F \ge F^\star$. Then the sparse-penalized DNN estimator*

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}_n^{\mathrm{DNN}}} \left[\frac{1}{n}\sum_{i=1}^n \ell(Y_i f(\mathbf{X}_i)) + J_{\lambda_n,\tau_n}(f)\right] \tag{2.3.16}$$

where

$$J_{\lambda_n,\tau_n}(f) := \lambda_n \left\|\boldsymbol{\theta}(f)\right\|_{\mathsf{clip},\tau_n}, \tag{2.3.17}$$

with $\lambda_n \gtrsim \log^3 n/n$ and $\tau_n \lesssim n^{-1}L_n^{-1}(N_nB_n)^{-L_n}$ satisfies

$$\mathsf{E}\left[\mathcal{E}_{\mathsf{P}}(\hat{f}_n)\right] \le 2 \inf_{f \in \mathcal{F}_n^{\mathrm{DNN}}} \left\{\mathcal{E}_{\mathsf{P}}(f) + J_{\lambda_n,\tau_n}(f)\right\} + c_3\frac{\log n}{n} \tag{2.3.18}$$

for some $c_3 > 0$, where the expectation is taken over the training data.

Proof. The proof is deferred to Section 2.7.2. □

We provide the theoretical guarantee for the performance of the sparse-penalized DNN estimator (2.3.16) with strictly convex margin-based losses in the binary classification task.

Theorem 2.3.4. *Let the loss function ℓ, the class of DNNs $\mathcal{F}_n^{\mathrm{DNN}} := \mathcal{F}_\rho^{\mathrm{DNN}}(L_n, N_n, B_n, F)$ and the penalty function $J_{\lambda_n,\tau_n}(\cdot)$ be as in Theorem 2.3.3. Let $\mathcal{F}^\star$ be a set of some real-valued functions on $[0,1]^d$. For $S > 0$, define $\mathcal{F}_{0,n}^{\mathrm{DNN}}(S)$ be as in Theorem 2.3.2. Assume that there are universal constants $\kappa > 0$, $r > 0$ and $c > 0$ such that*

$$\sup_{f^\diamond \in \mathcal{F}^\star} \inf_{f \in \mathcal{F}_{0,n}^{\mathrm{DNN}}(S)} \|f - f^\diamond\|_{2,\mathsf{P}_\mathsf{X}} \le c\left(\frac{S}{\log^r n}\right)^{-\kappa}. \tag{2.3.19}$$

for any $S > 0$ and $n \in \mathbb{N}$. Then the sparse-penalized DNN estimator defined by (2.3.16) satisfies

$$\sup_{\mathsf{P} \in \mathcal{Q}_{F^\star}: f_\ell^\star \in \mathcal{F}^\star} \mathsf{E}\left[\mathcal{E}_{\mathsf{P}}(\hat{f}_n)\right] \lesssim n^{-\frac{2\kappa}{2\kappa+1}} \log^{3+r} n. \tag{2.3.20}$$

Proof. The proof is deferred to Section 2.7.3. □

We may impose certain smoothness and/or structure to the optimal classification function as we did it for the regression function. Then we obtain the convergence rate of the excess risk by the above corollary. For example if the optimal classification function $f_l^\star$ is in a mixed smooth Besov space, the the excess risk of the sparse-penalized DNN estimator (2.3.16) with the ReLU activation function converges to zero with rate $n^{-\frac{2\alpha}{2\alpha+1}} \log^{3+d} n$.

2.4 Implementation

In this section, we propose a scalable optimization algorithm to solve the nonconvex problem (2.2.1), which is a combination of the concave-convex procedure (CCCP) [101] and the proximal gradient descent algorithm. For notational convenience, we let $f(\cdot|\boldsymbol{\theta})$ be a DNN having the parameter $\boldsymbol{\theta}$. Then we rewrite our objective function as

$$Q(\boldsymbol{\theta}) := \mathcal{L}_n(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_{\mathsf{Clip},\tau} \tag{2.4.1}$$

for given $\lambda > 0$ and $\tau > 0$, where

$$\mathcal{L}_n(\boldsymbol{\theta}) := n^{-1} \sum_{i=1}^{n} \ell(Y_i, f(\mathbf{X}_i|\boldsymbol{\theta}))$$

which denotes the empirical risk.

The CCCP algorithm is an optimization algorithm that is popularly used to minimize an objective function expressed as the sum of convex and concave functions. The CCCP updates the solution by iteratively minimizing the tight convex upper bound of the objective function at the current solution. The clipped L_1 penalty can be decomposed to convex and concave

parts as

$$\begin{aligned}\|\boldsymbol{\theta}\|_{\mathsf{Clip},\tau} &:= \sum_{j=1}^{p}\left(\frac{|\theta_j|}{\tau}\wedge 1\right)\\ &= \frac{1}{\tau}\sum_{j=1}^{p}|\theta_j| - \frac{1}{\tau}\sum_{j=1}^{p}(|\theta_j|-\tau)\mathbb{1}(|\theta_j|\geq\tau).\end{aligned}$$

where p is the dimension of $\boldsymbol{\theta}$ and the first term of the right-hand side is convex while the second term is concave in $\boldsymbol{\theta}$. For given current solution $\hat{\theta}^{(t)}$, the tight convex upper bound of the second term $-\frac{1}{\tau}\sum_{j=1}^{p}(|\theta_j|-\tau)I(|\theta_j|\geq\tau)$ at the current solution $\hat{\theta}_j^{(t)}$ is given as

$$-\frac{1}{\tau}\sum_{j=1}^{p}\operatorname{sign}\left(\hat{\theta}_j^{(t)}\right)\left(\theta_j+\tau\right)\mathbb{1}\left(|\hat{\theta}_j^{(t)}|>\tau\right).$$

Then the upper bound of the objective function (up to a constant) is given by

$$Q^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) := \mathcal{L}_n(\boldsymbol{\theta}) - \left\langle\frac{\lambda}{\tau}\mathbf{h}_\tau^{(t)},\boldsymbol{\theta}+\tau\mathbf{1}\right\rangle + \frac{\lambda}{\tau}\|\boldsymbol{\theta}\|_1, \tag{2.4.2}$$

where

$$\mathbf{h}_\tau^{(t)} := \left(\operatorname{sign}(\hat{\boldsymbol{\theta}}_j^{(t)})\mathbb{1}\left(|\hat{\boldsymbol{\theta}}_j^{(t)}|>\tau\right)\right)_{j\in[p]}.$$

The minimization of the upper bound $Q^*(\cdot|\boldsymbol{\theta}^{(t)})$ is justified by the following proposition.

Proposition 2.4.1. *If the parameter $\tilde{\boldsymbol{\theta}}$ satisfies $Q^*(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^{(t)}) \leq Q^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, then $Q(\tilde{\boldsymbol{\theta}}) \leq Q(\boldsymbol{\theta}^{(t)})$.*

Proof. By definition of $Q^*(\cdot|\boldsymbol{\theta}^{(t)})$, $Q^*(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}^{(t)})$ and $Q(\tilde{\boldsymbol{\theta}}) \leq Q^*(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^{(t)})$, which lead to the desired result. □

We apply the proximal gradient descent algorithm to do decrease the CCCP upper bound $Q^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in (2.4.2). That is, we update the solution as

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \mathsf{Prox}_{\eta_t(\frac{\lambda}{\tau}\|\cdot\|_1)}\left(\boldsymbol{\theta}^{(t)} - \eta_t \nabla Q^*(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})\right) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left[\frac{\lambda}{\tau}\|\boldsymbol{\theta}\|_1 + \left\langle \nabla \mathcal{L}_n(\boldsymbol{\theta}^{(t)}) - \frac{\lambda}{\tau}\mathbf{h}_\tau^{(t)}, \boldsymbol{\theta}\right\rangle + \frac{1}{2\eta_t}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\right\|_2^2\right]\end{aligned} \tag{2.4.3}$$

where η_t is a pre-specified learning rate and the proximal mapping is defined as $\mathsf{Prox}_{\eta g}(\boldsymbol{\theta}) := \operatorname{argmin}_{\tilde{\boldsymbol{\theta}}}\left[g(\tilde{\boldsymbol{\theta}}) + \frac{1}{2\eta}\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2\right]$.

The advantage of our optimization algorithm is that (2.4.3) has the exactly sparse solution given by

$$\hat{\theta}_j^{(t+1)} = \left(u_{\tau,\lambda,j}^{(t)} - \operatorname{sign}(u_{\tau,\lambda,j}^{(t)})\eta_t\frac{\lambda}{\tau}\right)\mathbb{1}\left(\left|u_{\tau,\lambda,j}^{(t)}\right| \geq \eta_t\frac{\lambda}{\tau}\right), \tag{2.4.4}$$

where

$$u_{\tau,\lambda,j}^{(t)} := \hat{\theta}_j^{(t)} - \eta_t\left(\nabla \mathcal{L}_n(\hat{\theta}_j^{(t)}) + \frac{\lambda}{\tau}h_{\tau,j}^{(t)}\right)$$

for $j \in [p]$. The solution (2.4.4) is a soft-thresholded version of $u_{\tau,\lambda,j}^{(t)}$, which can be sparse. Thus we can obtain a sparse estimates of the DNN parameter during the training procedure, without any post-training pruning algorithm such as [43].

2.5 Numerical studies

2.5.1 Regression with simulated data

In this section, we carry out simulation studies to illustrate the finite-sample performance of the proposed sparse-penalized DNN estimator (SDNN). We compare the proposed method with other popularly used regression estimators: kernel ridge regression (KRR), k-nearest neighbors (kNN), random forest (RF), and non-sparse DNN (NSDNN).

For kernel ridge regression, we used a radial basis function (RBF) kernel. For both the non-sparse and sparse DNN estimators, we used a network architecture of 5 hidden layers with the numbers of hidden nodes $(100, 100, 100, 100, 100)$. The non-sparse DNN is learned with popularly used optimizing algorithm Adam with learning rate 10^{-3}.

We select tuning parameters associated with each estimator by optimizing the performance on a held-out validation set whose size is one-fifth of the size of whole training data. The scale parameter of the RBF kernel and a degree of regularization for kernel ridge regression, the number of neighbors for k-nearest neighbors, and the depth of the trees for the random forest are chosen in this way. For the sparse-penalized DNN estimator, two tuning parameters λ and τ are selected among different combinations of tuning parameter values.

We first generate 10-dimensional input data points from the uniform distribution on $[0, 1]^{10}$. For each input $\mathbf{x}$, the corresponding response Y is generated as $Y = f^\star(\mathbf{x}) + \epsilon$ for some function $f^\star$, where ϵ is a standard normal error. The functions used as $f^\star$ for the generation of the simulation data are as listed below:

$$
\begin{aligned}
f_1^\star(\mathbf{x}) &= c_1 \sum_{j=1}^{10} (-1)^j j^{-1} x_j \\
f_2^\star(\mathbf{x}) &= c_2 \sin(\|\mathbf{x}\|_1) \\
f_3^\star(\mathbf{x}) &= c_3 \left[x_1 x_2^2 - x_3 + \log\left(x_4 + 4x_5 + \exp(x_6 x_7 - 5x_5)\right) + \tan(x_8 + 0.1) \right] \\
f_4^\star(\mathbf{x}) &= c_4 \left[\exp\left(3x_1 + x_2^2 - \sqrt{x_3 + 5}\right) + 0.01 \cot\left(\frac{1}{0.01 + |x_4 - 2x_5 + x_6|}\right) \right] \\
f_5^\star(\mathbf{x}) &= c_5 \left[3 \exp(\|\mathbf{x}\|_2) \mathbb{1}\left(x_2 \geq x_3^2\right) + x_3^{x_4} - x_5 x_6 x_7^4 \right] \\
f_6^\star(\mathbf{x}) &= c_6 \Big[4 x_1 x_2 x_3 x_4 \mathbb{1}\left(x_3 + x_4 \geq 1, x_5 \geq x_6\right) + \tan(\|\mathbf{x}\|_1) \mathbb{1}\left(x_1^2 x_7 x_8 \geq x_9 x_{10}^3\right) \Big].
\end{aligned}
$$

The functions $f_1^\star$ and $f_2^\star$ are globally smooth functions, f_3 and f_4 are composition structured functions and f_5 and f_6 are piecewise smooth functions. The constants $c_1, \ldots, c_6$ are chosen so that the error variance becomes 5% of the variance of the response. The variance due to the regression function is approximated by its empirical version based on 10^5 generated input values.

TABLE 2.1: Simulation results for $f_1^\star$ and $f_2^\star$. We show the average empirical L_2 error with standard deviation in parenthesis from 50 simulation replicates.

$f^\star$	$f_1^\star$		$f_2^\star$	
n	100	200	100	200
KRR	1.8746 (0.1105)	1.3127 (0.0959)	3.3593 (0.2094)	2.6436 (0.1004)
kNN	2.6146 (0.1533)	2.3013 (0.104)	3.5901 (0.2204)	3.2493 (0.1434)
RF	2.3027 (0.1451)	1.9293 (0.0847)	4.1124 (0.1464)	3.8575 (0.0676)
NSDNN	1.1261 (0.1275)	1.1345 (0.1149)	4.1484 (0.8753)	2.3146 (0.6926)
SDNN	**0.8267 (0.2202)**	**0.7439 (0.2176)**	**3.0595 (0.7483)**	**2.2139 (0.4261)**

TABLE 2.2: Simulation results for $f_3^\star$ and $f_4^\star$. We show the average empirical L_2 error with standard deviation in parenthesis from 50 simulation replicates.

$f^\star$	$f_3^\star$		$f_4^\star$	
n	100	200	100	200
KRR	2.2382 (0.1678)	1.6025 (0.084)	1.9906 (0.1167)	1.7481 (0.0731)
kNN	2.6611 (0.2103)	2.3431 (0.1274)	2.1893 (0.1217)	2.0328 (0.1131)
RF	2.4664 (0.1827)	2.0914 (0.1059)	1.7605 (0.134)	1.5364 (0.2021)
NSDNN	1.4774 (0.1522)	1.3078 (0.0908)	1.5445 (0.2487)	1.6622 (0.9068)
SDNN	**1.1675 (0.0692)**	**1.0353 (0.0651)**	**1.3392 (0.3541)**	**1.28 (0.8453)**

The performance of each estimator is measured by the empirical L_2 error computed based on newly generated 10^5 input values. The results are presented in Table 2.1, Table 2.2 and Table 2.3. We report the mean value of the empirical L_2 errors over 50 simulation replicates. We see that the proposed sparse-penalized DNN estimator outperforms the other competing estimators for both globally smooth, composition structured and piecewise smooth functions.

2.5.2 Classification with real data

We compare the proposed method with other competing methods in the following four data sets from the UCI repository:

- Haberman: Haberman's survival data set contains 306 patients who had undergone surgery for breast cancer at the University of Chicago's

TABLE 2.3: Simulation results for $f_5^\star$ and $f_6^\star$. We show the average empirical L_2 error with standard deviation in parenthesis from 50 simulation replicates.

$f^\star$	$f_5^\star$		$f_6^\star$	
n	100	200	100	200
KRR	3.2533 (0.0902)	2.9729 (0.0575)	3.1292 (0.1485)	2.8269 (0.0979)
kNN	3.4834 (0.1694)	3.2757 (0.13)	3.4473 (0.1442)	3.2318 (0.1071)
RF	3.0214 (0.1207)	2.6889 (0.092)	3.3471 (0.0889)	3.0714 (0.0721)
NSDNN	3.1779 (0.2885)	2.7072 (0.225)	3.2556 (0.251)	2.989 (0.1904)
SDNN	**2.5524 (0.1981)**	**2.1513 (0.1826)**	**2.8798 (0.1998)**	**2.6544 (0.1206)**

Billings Hospital. The task is to predict whether each patient survives after 5 years after the surgery or not.

- Retinopathy: This data set contains features extracted from 1,151 eye's images. The task is to predict whether an eye's image contains sings of diabetic retinopathy or not based on the other features.
- Tic-tac-toe: This data set contains all the 957 possible board configurations at the end of tic-tac-toe games, which are encoded to 27 input variables. The task is to predict the winner of the game.
- Promoter: This data set consists of A, C, G, T nucleotides at 57 positions for 106 gene sequnces. Each nucleotide is encoded to a 3-dimensional one-hot vector. The task is to predict whether a gene is promoters or non-promoter.

For competing methods, we considered a support vector machine (SVN), k-nearest neighbors (kNN), random forest (RF), and non-sparse DNN (NSDNN). For the support vector machine, we used the RBF kernel. The scale parameter of the kernel and a degree of regularization are selected by evaluation on a validation set. Setup for the other estimator is the same as the setup in Section 2.5.1.

We split the whole data into training and test data sets with the ratio 7:3, then evaluate the classification accuracy of each learned estimator on the test data set. We repeat this splits 50 times. Table 2.4 presents the averaged classification accuracy over 50 training-test splits. The proposed sparse-penalized DNN estimator performs best for Tic-tac-toe and Promoter data sets and performs second best for the other two data sets.

TABLE 2.4: Simulation results with UCI data sets. We show the average classification accuracy with standard deviation in parenthesis from 50 training-test splits.

Data	Haberman	Retinopathy	Tic-tac-toe	Promoter
(n, d)	(214, 3)	(805, 19)	(669, 27)	(74, 171)
SVM	0.7298 (0.0367)	0.5737 (0.0282)	0.8467 (0.0243)	0.7887 (0.1041)
kNN	**0.7587 (0.0366)**	0.6436 (0.0263)	0.9714 (0.0102)	0.8012 (0.0649)
RF	0.7365 (0.0377)	0.665 (0.0263)	0.9777 (0.0103)	0.8725 (0.0582)
NSDNN	0.7328 (0.0464)	**0.7158 (0.0293)**	0.9735 (0.0107)	0.8594 (0.062)
SDNN	0.752 (0.0382)	0.6987 (0.0375)	**0.98 (0.0085)**	**0.8769 (0.0474)**

2.6 Conclusion

In this chapter, we proposed a novel sparse-penalized DNN estimator leaned with the clipped L_1 penalty. The continuity of the clipped L_1 penalty makes the optimization tractable, but it is challenging due to the nonconvexity of the penalty. We proposed the efficient and scalable optimization algorithm for this nonconvex optimization. The proposed optimization algorithm monotonically decrease the objective function at every iteration. Theoretically, we showed that the proposed sparse-penalized DNN estimator attains minimax optimality adaptively for several regression problems considered in related literature. Also, we derived convergence rates of the sparse-penalized DNN estimator learned with strictly convex loss functions for binary classification.

For the binary classification, we only consider the strictly convex losses which are popular for learning DNNs. Using the hinge loss is promising since it can lead to the optimal DNN classifier in several true probability models (in which learning with the logistic loss may yield suboptimal classifiers) [51]. We expect that the sparse-penalized DNN estimator learned with the hinge loss and the clipped L_1 penalty can attain the minimax optimal convergence rates for such true probability models. We will investigate this issue shortly.

2.7 Proofs

2.7.1 Covering numbers of classes of DNNs

We provide covering number bounds for classes of DNNs.

The covering number of the function class is defined as follows. Let $\mathcal{F}$ be a given class of real-valued functions defined on $\mathcal{X}$. Let $\delta > 0$. A collection $\{f_i : i \in [N]\}$ is called a *δ-covering set* of $\mathcal{F}$ with respect to the norm $\|\cdot\|$ if, for all $f \in \mathcal{F}$, there exists f_i in the collection such that $\|f - f_i\| \leq \delta$. The cardinality of the minimal δ-covering set is called the *δ-covering number* of $\mathcal{F}$ with respect to the norm $\|\cdot\|$, and is denoted by $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|)$.

Proposition 2.7.1 (Lemma 3 of [87], Proposition 1 of [71])**.** *Let $L \in \mathbb{N}$, $N \in \mathbb{N} \setminus \{1\}$, $B \geq 1$, $F > 0$ and ρ the 1-Lipschitz activation function. Let*

$$\mathcal{F}_{\rho,0}^{\mathrm{DNN}}(S) := \left\{ f \in \mathcal{F}_{\rho}^{\mathrm{DNN}}(L, N, B, F) : \|\boldsymbol{\theta}(f)\|_0 \leq S \right\}.$$

Then we have that for any $\delta > 0$,

$$\log \mathcal{N}\left(\delta, \mathcal{F}_{\rho,0}^{\mathrm{DNN}}(S), \|\cdot\|_\infty\right) \leq 2S(L+1) \log\left(\frac{LBN}{\delta}\right) \tag{2.7.1}$$

We can compute $\delta-$covering number of the class of DNNs with a restriction on the clipped L_1 norm, for δ not too small, by the following proposition.

Proposition 2.7.2. *Let $L \in \mathbb{N}$, $N \in \mathbb{N} \setminus \{1\}$, $B \geq 1$, $F > 0$, τ and ρ the 1-Lipschitz activation function. Let*

$$\mathcal{F}_{\rho,\tau}^{\mathrm{DNN}}(S) := \left\{ f \in \mathcal{F}_{\rho}^{\mathrm{DNN}}(L, N, B, F) : \|\boldsymbol{\theta}(f)\|_{\mathsf{clip},\tau} \leq S \right\}.$$

Then we have that for any $\delta > \tau L(BN)^L$,

$$\begin{aligned} \mathcal{N}\left(\delta, \mathcal{F}_{\rho,\tau}^{\mathrm{DNN}}(S), \|\cdot\|_\infty\right) &\leq \mathcal{N}\left(\delta - \tau L(BN)^L, \mathcal{F}_{\rho,0}^{\mathrm{DNN}}(S), \|\cdot\|_\infty\right) \\ &\leq 2S(L+1) \log\left(\frac{LBN}{\delta - \tau L(BN)^L}\right) \end{aligned} \tag{2.7.2}$$

Proof. For a DNN f with parameter $\boldsymbol{\theta}(f)$, we let $f^{(\tau)}$ be the DNN constructed by the parameter which is the hard thresholding of $\boldsymbol{\theta}(f)$ with the threshold τ, that is, $\boldsymbol{\theta}(f^{(\tau)}) = \boldsymbol{\theta}(f)\mathbb{1}(|\boldsymbol{\theta}(f)| > \tau)$. Then by Lemma 2.7.3,

$$\|f - f^{(\tau)}\|_\infty \le L(BN)^L \|\boldsymbol{\theta}(f) - \boldsymbol{\theta}(f^{(\tau)})\|_\infty \le \tau L(BN)^L.$$

Given $\delta > 2\tau L(BN)^L$, let $\delta^* := \delta - \tau L(BN)^L > 0$ and let $\{f_j^0 : j \in [N_{\delta^*}]\}$ be the minimal δ^*-covering set of $\mathcal{F}_{\rho,0}^{\mathrm{DNN}}(S)$ with respect to the norm $\|\cdot\|_\infty$, where $N_{\delta^*} := \mathcal{N}(\delta^*, \mathcal{F}_{\rho,0}^{\mathrm{DNN}}(S), \|\cdot\|_\infty)$. Since $\|\boldsymbol{\theta}(f^{(\tau)})\|_0 = \|\boldsymbol{\theta}(f^{(\tau)})\|_{\mathsf{clip},\tau} \le \|\boldsymbol{\theta}(f)\|_{\mathsf{clip},\tau} \le S$, it follows that $f^{(\tau)} \in \mathcal{F}_{\rho,0}^{\mathrm{DNN}}(S)$ for any $f \in \mathcal{F}_{\rho,\tau}^{\mathrm{DNN}}(S)$. Hence for any $f \in \mathcal{F}_{\rho,\tau}^{\mathrm{DNN}}(S)$, there is $j \in [N_{\delta^*}]$ such that $\|f^{(\tau)} - f_j^0\|_\infty \le \delta^*$ and so

$$\begin{aligned}\|f - f_j^0\|_\infty &\le \|f - f^{(\tau)}\|_\infty + \|f^{(\tau)} - f_j^0\|_\infty \\ &\le \tau L(BN)^L + \delta^* = \delta,\end{aligned}$$

which implies that $\{f_j^0 : j \in [N_{\delta^*}]\}$ is the δ-covering set of $\mathcal{F}_{\rho,\tau}^{\mathrm{DNN}}(S)$. □

Lemma 2.7.3. *Let $L \in \mathbb{N}$, $N \in \mathbb{N} \setminus \{1\}$, $B \ge 1$, and ρ the 1-Lipschitz activation function. For any two DNNs $f_1, f_2 \in \mathcal{F}_\rho(L, N, B, \infty)$, we have*

$$\|f_1 - f_2\|_{\infty,[0,1]^d} \le L(BN)^L \|\boldsymbol{\theta}(f_1) - \boldsymbol{\theta}(f_2)\|_\infty$$

Proof. For $f \in \mathcal{F}_\rho(L, N, B, \infty)$ expressed as

$$f(\mathbf{x}) = A_{L+1} \circ \rho_L \circ A_L \circ \cdots \circ \rho_1 \circ A_1(\mathbf{x}),$$

we define $[f]_l^- : [0,1]^d \mapsto \mathbb{R}^{N-1}$ and $[f]_l^+ : \mathbb{R}^{N-1} \mapsto \mathbb{R}$ for $l \in \{2, \ldots, L\}$ by

$$\begin{aligned}[f]_l^-(\cdot) &:= \rho_{l-1} \circ A_{l-1} \circ \cdots \circ \rho_1 \circ A_1(\cdot), \\ [f]_l^+(\cdot) &:= A_{L+1} \circ \rho_L \circ A_L \circ \cdots \rho_l \circ \mathsf{A}_l \circ \rho_{l-1}(\cdot).\end{aligned}$$

Corresponding to the last and first layer, we define $f_1^-(\mathbf{x}) = \mathbf{x}$ and $f_{L+1}^+(\mathbf{x}) = \mathbf{x}$. Note that $f = [f]_{l+1}^+ \circ A_l \circ [f]_l^-$.

Let $\mathbf{W}_l$ and $\mathbf{b}_l$ be the weight matrix and bias vector at the l-th hidden layer of f. Note that both the numbers of rows and columns of $\mathbf{W}_l$ are less

than $N-1$. Then for any $\mathbf{x} \in [0,1]^d$

$$\begin{aligned}
\left\| [f]_l^-(\mathbf{x}) \right\|_\infty = \left\| \mathbf{W}_l [f]_{l-1}^-(\mathbf{x}) + \mathbf{b}_l \right\|_\infty &\le (N-1)B \left\| [f]_{l-1}^-(\mathbf{x}) \right\|_\infty + B \\
&\le NB \left(\left\| [f]_{l-1}^-(\mathbf{x}) \right\|_\infty \vee 1 \right) \\
&\le NB \left((NB \left\| [f]_{l-2}^-(\mathbf{x}) \right\|_\infty \vee 1) \vee 1 \right) \\
&\le (NB)^2 \left(\left\| [f]_{l-2}^-(\mathbf{x}) \right\|_\infty \vee 1 \right) \\
&\le (NB)^{l-1} \left(\|\mathbf{x}\|_\infty \vee 1 \right) = (NB)^{l-1},
\end{aligned}$$

where the fourth inequality follows from the assumption that $NB \ge 1$. Similarly, for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{N-1}$,

$$\left| [f]_{l+1}^+(\mathbf{z}_1) - [f]_{l+1}^+(\mathbf{z}_2) \right| \le (NB)^{L-l} \|\mathbf{z}_1 - \mathbf{z}_2\|_\infty$$

For $f_1, f_2 \in \mathcal{F}_\rho(L, N, B, \infty)$, letting $A_{j,l}$ be the affine transform at the l-th hidden layer of f_j for $j = 1, 2$, we have for any $\mathbf{x} \in [0,1]^d$,

$$\begin{aligned}
|f_1(\mathbf{x}) - f_2(\mathbf{x})| &\le \left| \sum_{l=1}^L \left[[f_1]_{l+1}^+ \circ A_{1,l} \circ [f_2]_l^-(\mathbf{x}) - [f_1]_{l+1}^+ \circ A_{2,l} \circ [f_2]_l^-(\mathbf{x}) \right] \right| \\
&\le \sum_{l=1}^L (BN)^{L-l} \left\| (A_{1,l} - A_{2,l}) \circ [f_2]_l^-(\mathbf{x}) \right\|_\infty \\
&\le \sum_{l=1}^L (BN)^{L-l} \|\boldsymbol{\theta}(f_1) - \boldsymbol{\theta}(f_2)\|_\infty \left\{ (N-1) \left\| [f_2]_l^-(\mathbf{x}) \right\|_\infty + 1 \right\} \\
&\le \sum_{l=1}^L (BN)^{L-l} \|\boldsymbol{\theta}(f_1) - \boldsymbol{\theta}(f_2)\|_\infty \left\{ (N-1)(NB)^{l-1} + 1 \right\} \\
&\le L(BN)^L \|\boldsymbol{\theta}(f_1) - \boldsymbol{\theta}(f_2)\|_\infty,
\end{aligned}$$

which completes the proof. □

2.7.2 Proofs of Theorem 2.3.1 and Theorem 2.3.3

Let P_n be the empirical distribution based on the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. We use the abbreviation $\mathsf{Q}f := \int f \mathrm{d}\mathsf{Q}$ for a measurable function f and measure Q.

For the proofs of Theorem 2.3.1 and Theorem 2.3.3, we need the following large deviation bound for empirical processes. This is a slight modification of Theorem 19.3 of [40] which states the result with the covering numbers with respect to the empirical L_2 norm. But since the empirical L_2 norm is always less than the L_∞ norm, the following lemma is a direct consequence of Theorem 19.3 of [40].

Lemma 2.7.4 (Theorem 19.3 of [40]). *Let $K_1 \geq 1$ and $K_2 \geq 1$. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be independent and identically distributed random variables with values in $\mathcal{Z}$ and $\mathcal{G}$ be a class of functions $g : \mathcal{Z} \mapsto \mathbb{R}$ with the properties $\|g\|_\infty \leq K_1$ and $\mathsf{E}g(\mathbf{Z})^2 \leq K_2 \mathsf{E}g(\mathbf{Z})$. Let $\omega \in (0,1)$ and $t^* > 0$. Assume that*

$$\sqrt{n}\omega\sqrt{1-\omega}\sqrt{t^*} \geq 288 \max\{2K_1, \sqrt{2K_2}\} \tag{2.7.3}$$

and that any $\delta \geq t^*/8$,

$$\frac{\sqrt{n}\omega(1-\omega)\delta}{96\sqrt{2}\max\{K_1, 2K_2\}} \geq \int_{\frac{\omega(1-\omega)\delta}{16\max\{K_1,2K_2\}}}^{\sqrt{\delta}} \sqrt{\log \mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_\infty\right)} \mathrm{d}u. \tag{2.7.4}$$

Then

$$\mathsf{P}\left(\sup_{g\in\mathcal{G}} \frac{\left|(\mathsf{P}-\mathsf{P}_n)g\right|}{t^* + \mathsf{P}g} \geq \omega\right) \leq 60 \exp\left(-\frac{nt^*\omega^2(1-\omega)}{128 \cdot 2304 \max\{K_1^2, K_2\}}\right).$$

Proof of Theorem 2.3.1. Let $K_n := (\sqrt{4/a}\log^{1/2} n) \vee F$. Let

$$Y^\dagger := \mathrm{sign}(Y)(|Y| \wedge K_n),$$

which is the truncated version of Y and $f^\dagger$ be the regression function of $Y^\dagger$, that is,

$$f^\dagger(\mathbf{x}) := \mathsf{E}(Y^\dagger | \mathbf{X} = \mathbf{x}).$$

We suppress the dependency on n in the notation $Y^\dagger$ and $f^\dagger$ for simple presentation. We start with the decomposition

$$\|\hat{f}_n - f^*\|_{2,\mathsf{P}}^2 = \mathsf{P}(Y - \hat{f}_n(\mathbf{X}))^2 - \mathsf{P}(Y - f^\star(\mathbf{X}))^2 = \sum_{i=1}^{4} A_{i,n}, \tag{2.7.5}$$

where

$$\begin{aligned}
A_{1,n} &:= \left[\mathsf{P}(Y - \hat{f}_n(\mathbf{X}))^2 - \mathsf{P}(Y - f^\star(\mathbf{X}))^2\right] \\
&\quad - \left[\mathsf{P}(Y^\dagger - \hat{f}_n(\mathbf{X}))^2 - \mathsf{P}(Y^\dagger - f^\dagger(\mathbf{X}))^2\right] \\
A_{2,n} &:= \left[\mathsf{P}(Y^\dagger - \hat{f}_n(\mathbf{X}))^2 - \mathsf{P}(Y^\dagger - f^\dagger(\mathbf{X}))^2\right] \\
&\quad - 2\left[\mathsf{P}_n(Y^\dagger - \hat{f}_n(\mathbf{X}))^2 - \mathsf{P}_n(Y^\dagger - f^\dagger(\mathbf{X}))^2\right] - 2J_{\lambda_n,\tau_n}(\hat{f}_n) \\
A_{3,n} &:= 2\left[\mathsf{P}_n(Y^\dagger - \hat{f}(\mathbf{X}))^2 - \mathsf{P}_n(Y^\dagger - f^\dagger(\mathbf{X}))^2\right] \\
&\quad - 2\left[\mathsf{P}_n(Y - \hat{f}(\mathbf{X}))^2 - \mathsf{P}_n(Y - f^\star(\mathbf{X}))^2\right] \\
A_{4,n} &:= 2\left[\mathsf{P}_n(Y - \hat{f}(\mathbf{X}))^2 - \mathsf{P}_n(Y - f^\star(\mathbf{X}))^2\right] + 2J_{\lambda_n,\tau_n}(\hat{f}_n)
\end{aligned}$$

For $A_{1,n}$, we let

$$\begin{aligned}
A_{1,1,n} &:= \mathsf{P}\left((Y^\dagger - Y)(2\hat{f}_n(\mathbf{X}) - Y - Y^\dagger)\right) \\
A_{1,2,n} &:= \mathsf{P}\left\{\left(Y^\dagger - f^\dagger(\mathbf{X})) - Y + f^\star(\mathbf{X})\right)\left(Y^\dagger - f^\dagger(\mathbf{X}) + Y - f^\star(\mathbf{X})\right)\right\}
\end{aligned}$$

so that $A_{1,n} = A_{1,1,n} + A_{1,2,n}$. We use the Cauchy-Schwarz inequality to get

$$|A_{1,1,n}| \le \sqrt{\mathsf{P}\left(Y^\dagger - Y\right)^2}\sqrt{\mathsf{P}\left(2\hat{f}_n(\mathbf{X}) - Y - Y^\dagger\right)^2}.$$

We will bound each term of the preceding display. For the first term, by the the assumption that $\mathsf{E}e^{aY^2/2} \le e^{a(F^\star)^2}\mathsf{E}e^{a\epsilon^2} \le c_{11} < \infty$ for some $c_{11} > 0$, we

have

$$\begin{aligned}\mathsf{P}\left(Y^\dagger - Y\right)^2 &= \mathsf{P}[|Y|^2 \mathbb{1}(|Y| > K_n)] \\ &\leq \mathsf{P}[4a^{-1}\mathrm{e}^{aY^2/4}\mathrm{e}^{aY^2/4 - aK_n^2/4}] \\ &\leq 4a^{-1}c_{11}\mathrm{e}^{-aK_n^2/4} \lesssim n^{-1}.\end{aligned}$$

For the second term, we have

$$\begin{aligned}\mathsf{P}\left(2\hat{f}_n(\mathbf{X}) - Y - Y^\dagger\right)^2 &\leq 2\mathsf{P}(Y^2) + 2\mathsf{P}(2\hat{f}_n(\mathbf{X}) - Y^\dagger)^2 \\ &\leq 2a^{-1}\mathsf{P}(\mathrm{e}^{aY^2}) + 18K_n^2 \leq c_{12}\log n\end{aligned}$$

for some $c_{12} > 0$. So $|A_{1,1,n}| \lesssim \log n/n$. For $A_{1,2,n}$, using the Cauchy-Schwarz inequality we have

$$\begin{aligned}|A_{1,2,n}| \leq &\sqrt{2\mathsf{P}(Y^\dagger - Y)^2 + 2\mathsf{P}(f^\dagger(\mathbf{X}) - f^\star(\mathbf{X}))^2} \\ &\times \sqrt{2\mathsf{P}(Y^2) + 2\mathsf{P}(Y^\dagger - f^\dagger(\mathbf{X}) - f^\star(\mathbf{X}))^2}.\end{aligned}$$

Using the similar arguments as before, we can bound the terms $\mathsf{P}(Y^\dagger - Y)^2$ and $\mathsf{P}(Y^2)$. Then since $\|f^\star\|_\infty \leq F^\star$, we have $|A_{1,2,n}| \lesssim \log n/n$.

The term $\mathsf{E}(A_{3,n})$ is bounded above by $\log n/n$ up to a constant by the similar way of bounding $A_{1,n}$.

For $A_{2,n}$, we let $\Delta(f)(\mathbf{Z}) := (Y^\dagger - f(\mathbf{X}))^2 - (Y^\dagger - f^\dagger(\mathbf{X}))^2$ with $\mathbf{Z} := (\mathbf{X}, Y)$ for $f \in \mathcal{F}$ for simplicity. For $t > 0$, we can write

$$\begin{aligned}\mathsf{P}(A_{2,n} > t) &\leq \mathsf{P}\left(\sup_{f \in \mathcal{F}_n^{\mathrm{DNN}}} \frac{(\mathsf{P} - \mathsf{P}_n)\Delta(f)}{t + 2J_{\lambda_n,\tau_n}(f) + \mathsf{P}\Delta(f)} \geq \frac{1}{2}\right) \\ &\leq \sum_{j=0}^{\infty} \mathsf{P}\left(\sup_{f \in \mathcal{F}_{n,j,t}} \frac{(\mathsf{P} - \mathsf{P}_n)\Delta(f)}{2^j t + \mathsf{P}\Delta(f)} \geq \frac{1}{2}\right)\end{aligned}$$

where we define

$$\mathcal{F}_{n,j,t} := \left\{f \in \mathcal{F}_n^{\mathrm{DNN}} : 2^{j-1}\mathbb{1}(j \neq 0)t \leq J_{\lambda_n,\tau_n}(f) \leq 2^j t\right\}.$$

We now apply Lemma 2.7.4 to the class of functions

$$\mathcal{G}_{n,j,t} := \left\{ \Delta(f) : [0,1]^d \times \mathbb{R} \mapsto \mathbb{R} : f \in \mathcal{F}_{n,j,t} \right\}.$$

We will check the conditions of Lemma 2.7.4. First for sufficiently large n, we have that for every $g \in \mathcal{G}_{n,j,t}$, $\|g\|_\infty \le 8K_n^2$ and

$$\begin{aligned} \mathsf{P}(g(\mathbf{Z}))^2 &= \mathsf{P}(Y^\dagger - f(\mathbf{X}) - (Y^\dagger - f^\dagger(\mathbf{X})))^2 (Y^\dagger - f(\mathbf{X}) + (Y^\dagger - f^\dagger(\mathbf{X})))^2 \\ &\le 4(K_n + F)^2 \mathsf{P}(f(\mathbf{X}) - f^\dagger(\mathbf{X}))^2 \le 16 K_n^2 \mathsf{P} g(\mathbf{Z}) \end{aligned}$$

The condition (2.7.3) holds for any sufficiently large n if $t \gtrsim \log^2 n/n$. For the condition (2.7.4), we observe that

$$\begin{aligned} &\left| (y^\dagger - f_1(\mathbf{x}))^2 - (y^\dagger - f^\dagger(\mathbf{x}))^2 - \left\{ (y^\dagger - f_2(\mathbf{x}))^2 - (y^\dagger - f^\dagger(\mathbf{x}))^2 \right\} \right| \\ &\le |f_1(\mathbf{x}) - f_2(\mathbf{x})||f_1(\mathbf{x}) + f_2(\mathbf{x}) - 2y^\dagger| \\ &\le 4K_n |f_1(\mathbf{x}) - f_2(\mathbf{x})| \end{aligned}$$

for any $f_1, f_2 \in \mathcal{F}_{n,j,t}$ and $(\mathbf{x}, y) \in [0,1]^d \times \mathbb{R}$ and so

$$\mathcal{N}\left(u, \mathcal{G}_{n,j,t}, \|\cdot\|_\infty\right) \le \mathcal{N}\left(u/(4K_n), \mathcal{F}_{n,j,t}, \|\cdot\|_\infty\right).$$

With $\omega = 1/2$, by Proposition 2.7.2 and the assumption $\tau_n L_n (B_n N_n)^{L_n} \lesssim n^{-1}$, we have for any $\delta \ge c_{21} \log^2 n/n$ for some $c_{21} > 0$

$$\begin{aligned} &\int_{\delta/(2^{11}K_n^2)}^{\sqrt{\delta}} \log^{1/2} \mathcal{N}\left(\frac{u}{4K_n}, \mathcal{F}_{n,j,t}, \|\cdot\|_\infty\right) \mathrm{d}u \\ &\le \sqrt{\delta} \log^{1/2} \mathcal{N}\left(\frac{\delta}{2^{13}K_n^3}, \mathcal{F}_{n,j,t}, \|\cdot\|_\infty\right) \\ &\le \sqrt{\delta} (2^j t/\lambda_n)^{1/2} L_n^{1/2} \log^{1/2}\left(\frac{L_n B_n N_n}{\delta/(2^{13}K_n^3) - \tau_n L_n (B_n N_n)^{L_n}}\right) \\ &\le c_{22}\sqrt{\delta}\sqrt{2^j t} \log n/\sqrt{\lambda_n} \\ &\le c_{23}\sqrt{\delta}\sqrt{2^j t}\sqrt{n}/\log^{3/2} n \end{aligned} \tag{2.7.6}$$

for some $c_{22} > 0$ and $c_{23} > 0$. Thus the condition (2.7.4) is met if $\sqrt{\delta}/\log n \ge$

$c_{24}\sqrt{t}/\log^{3/2} n$ for any $\delta \geq 2^j t/8 \geq c_{21}\log^2 n/n$ where c_{24} is an universal constant, and this holds for sufficiently large n and $t \geq t_n := c_{25}\log^2 n/n$ for some constant $c_{25} > 0$. Then we have

$$\mathsf{P}(A_{2,n} > t) \lesssim \sum_{j=0}^{\infty} \exp\left(-c_{26}2^j \frac{nt}{\log n}\right) \lesssim \exp\left(-c_{26}\frac{nt}{\log n}\right),$$

which implies

$$\begin{aligned}\mathsf{E}(A_{2,n}) &\leq 2t_n + \int_{2t_n}^{\infty} \mathsf{P}(A_{2,n} > t)\mathrm{d}t \\ &\lesssim \frac{\log^2 n}{n} + \frac{\log^2 n}{n}\mathrm{e}^{-c_{27}\log n} \lesssim \frac{\log^2 n}{n}\end{aligned}$$

for some $c_{26}, c_{27} > 0$.

For $A_{4,n}$, we choose a neural network function $f_n^\circ \in \mathcal{F}_n^{\mathrm{DNN}}$ such that

$$\left\|f_n^\circ - f^\star\right\|_{2,\mathsf{P}_\mathbf{X}}^2 + J_{\lambda_n,\tau_n}(f_n^\circ) \leq \inf_{f\in\mathcal{F}_n^{\mathrm{DNN}}}\left[\left\|f - f^\star\right\|_{2,\mathsf{P}_\mathbf{X}}^2 + J_{\lambda_n,\tau_n}(f)\right] + n^{-1}$$

Then by the basic inequality $\mathsf{P}_n(Y - \hat{f}_n)^2 + J_{\lambda_n,\tau_n}(\hat{f}_n) \leq \mathsf{P}_n(Y - f)^2 + J_{\lambda_n,\tau_n}(f)$ for any $f \in \mathcal{F}_n^{\mathrm{DNN}}$, we have

$$\begin{aligned}A_{4,n} &\leq 2\left[\mathsf{P}_n(Y - \hat{f}_n(\mathbf{X}))^2 - \mathsf{P}_n(Y - f_n^\circ(\mathbf{X}))^2\right] + 2J_{\lambda_n,\tau_n}(\hat{f}_n) \\ &\quad + 2\left[\mathsf{P}_n(Y - f_n^\circ(\mathbf{X}))^2 - \mathsf{P}_n(Y - f^\star(\mathbf{X}))^2\right] \\ &\leq 2J_{\lambda_n,\tau_n}(f_n^\circ) + 2\left[\mathsf{P}_n(Y - f_n^\circ(\mathbf{X}))^2 - \mathsf{P}_n(Y - f^\star(\mathbf{X}))^2\right]\end{aligned}$$

and so

$$\begin{aligned}\mathsf{E}(A_{4,n}) &\leq 2J_{\lambda_n,\tau_n}(f_n^\circ) + 2\left\|f_n^\circ - f^\star\right\|_{2,\mathsf{P}_\mathbf{X}}^2 \\ &\leq 2\inf_{f\in\mathcal{F}_n^{\mathrm{DNN}}}\left[\left\|f - f^\star\right\|_{2,\mathsf{P}_\mathbf{X}}^2 + J_{\lambda_n,\tau_n}(f)\right] + \frac{1}{n}.\end{aligned}$$

Combining all the bounds we have derived, we get the desired result. □

Proof of Theorem 2.3.3. Since $\mathcal{F}_n^{\mathrm{DNN}}$ is uniformly bounded and ℓ is continuously differentiable, ℓ is locally Lipschitz. That is, there is a constant $c_1 > 0$

such that

$$|\ell(z_1) - \ell(z_2)| \le c_1|z_1 - z_2| \tag{2.7.7}$$

for any $z_1, z_2 \in [-F, F]$. On the other hand, since $F \ge F^\star$, there is a constant $c_2 > 0$ such that

$$\mathsf{E}(\ell(Yf(\mathbf{X})) - \ell(Yf_\ell^\star(\mathbf{X}))^2 \le c_2\mathsf{E}(\ell(Yf(\mathbf{X})) - \ell(Yf_\ell^\star(\mathbf{X})) \tag{2.7.8}$$

for any $f \in \{f \in \mathcal{F} : \|f\|_\infty \le F\}$. This is a well known fact about the strictly convex losses and the proof can be found in [74] (see Lemma 6.1 there).

We decompose $\mathcal{E}_\mathsf{P}(\hat{f}_n)$ as

$$\mathcal{E}_\mathsf{P}(\hat{f}_n) = \mathsf{P}\ell(Y\hat{f}_n(\mathbf{X})) - \mathsf{P}\ell(Yf^\star(\mathbf{X})) = B_{1,n} + B_{2,n},$$

where

$$\begin{aligned} B_{1,n} &:= \left[\mathsf{P}\ell(Y\hat{f}_n(\mathbf{X})) - \mathsf{P}\ell(Yf^\star(\mathbf{X}))\right] \\ &\quad - 2\left[\mathsf{P}_n\ell(Y\hat{f}_n(\mathbf{X})) - \mathsf{P}_n\ell(Yf^\star(\mathbf{X}))\right] - 2J_{\lambda_n,\tau_n}(\hat{f}_n) \\ B_{2,n} &:= 2\left[\mathsf{P}_n\ell(Y\hat{f}_n(\mathbf{X})) - \mathsf{P}_n\ell(Yf^\star(\mathbf{X}))\right] + 2J_{\lambda_n,\tau_n}(\hat{f}_n). \end{aligned}$$

We bound $B_{1,n}$ by using the similar argument as in the proof of Theorem 2.3.1. Let $\Delta(f)(\mathbf{Z}) := \ell(Yf(\mathbf{X})) - \ell(Yf^\star(\mathbf{X}))$ with $\mathbf{Z} := (\mathbf{X}, Y)$ and let

$$\mathcal{F}_{n,j,t} := \left\{f \in \mathcal{F}_n^{\mathrm{DNN}} : 2^{j-1}\mathbb{1}(j \ne 0)t \le J_{\lambda_n,\tau_n}(f) \le 2^j t\right\}.$$

Then for $t > 0$, we can write

$$\begin{aligned} \mathsf{P}(B_{1,n} > t) &\le \mathsf{P}\left(\sup_{f\in\mathcal{F}_n^{\mathrm{DNN}}} \frac{(\mathsf{P}-\mathsf{P}_n)\Delta(f)}{t + 2J_{\lambda_n,\tau_n}(f) + \mathsf{P}\Delta(f)} \ge \frac{1}{2}\right) \\ &\le \sum_{j=0}^{\infty} \mathsf{P}\left(\sup_{f\in\mathcal{F}_{n,j,t}} \frac{(\mathsf{P}-\mathsf{P}_n)\Delta(f)}{2^j t + \mathsf{P}\Delta(f)} \ge \frac{1}{2}\right). \end{aligned}$$

We now apply Lemma 2.7.4 to the class of functions

$$\mathcal{G}_{n,j,t} := \left\{\Delta(f) : [0,1]^d \times \{-1,1\} \mapsto \mathbb{R} : f \in \mathcal{F}_{n,j,t}\right\},$$

By the conditions (2.7.7) and (2.7.8), we can set $K_1 = 2c_1F$ and $K_2 = c_2$ in Lemma 2.7.4. The condition (2.7.3) holds for any sufficiently large n if $t \gtrsim \log n/n$. For the condition (2.7.4), we let $K' := K_1 \vee 2K_2 = (2c_1F) \vee 2c_2$ for simplicity. Then since ℓ is Lipschitz, using a similar argument of (2.7.6) in the proof of Theorem 2.3.1, we have that any $\delta \geq c_{11} \log n/n$ for some $c_{11} > 0$

$$\begin{aligned}
&\int_{\delta/(4K')}^{\sqrt{\delta}} \log^{1/2} \mathcal{N}\left(u, \mathcal{G}_{n,j,t}, \|\cdot\|_\infty\right) \mathrm{d}u \\
&\leq \int_{\delta/(4K')}^{\sqrt{\delta}} \log^{1/2} \mathcal{N}\left(u/c_2, \mathcal{G}_{n,j,t}, \|\cdot\|_\infty\right) \mathrm{d}u \\
&\leq \sqrt{\delta}(2^j t/\lambda_n)^{1/2} L_n^{1/2} \log^{1/2}\left(\frac{L_n B_n N_n}{\delta/(4c_2K') - \tau_n L_n (B_n N_n)^{L_n}}\right) \\
&\leq c_{12}\sqrt{\delta}\sqrt{2^j t}\log n/\sqrt{\lambda_n} \leq c_{13}\sqrt{\delta}\sqrt{2^j t}\sqrt{n}/\log^{1/2} n
\end{aligned}$$

for some $c_{12}, c_{13} > 0$, provided the assumption $\tau_n L_n (B_n N_n)^{L_n} \lesssim n^{-1}$. Therefore, the condition (2.7.4) is met if $\sqrt{\delta} \geq c_{11}\sqrt{t}/\log^{1/2} n$ for any $\delta \geq 2^j t/8 \geq c_{14} \log n/n$ where c_{14} is an universal constant, and this holds for sufficiently large n and $t \geq t_n := c_{15} \log n/n$ for some constant $c_{15} > 0$. Then we have

$$\begin{aligned}
\mathsf{E}(B_{1,n}) &\leq 2t_n + \int_{2t_n}^{\infty} \mathsf{P}(B_{1,n} > t)\mathrm{d}t \\
&\leq 2t_n + \int_{2t_n}^{\infty} \exp(-c_{16} n t)\mathrm{d}t \\
&\lesssim \frac{\log n}{n} + \frac{1}{n}\mathrm{e}^{-c_{17}\log n} \lesssim \frac{\log n}{n}
\end{aligned}$$

for some $c_{16}, c_{17} > 0$.

For $B_{2,n}$, we choose a neural network function $f_n^\circ \in \mathcal{F}_n^{\mathrm{DNN}}$ such that

$$\mathcal{E}_{\mathsf{P}}(f_n^\circ) + J_{\lambda_n,\tau_n}(f_n^\circ) \leq \inf_{f \in \mathcal{F}_n^{\mathrm{DNN}}} \left[\mathcal{E}_{\mathsf{P}}(f) + J_{\lambda_n,\tau_n}(f)\right] + n^{-1}.$$

Then by the basic inequality $\mathsf{P}_n\ell(Y\hat{f}_n(\mathbf{X})) + J_{\lambda_n,\tau_n}(\hat{f}_n) \le \mathsf{P}_n\ell(Yf(\mathbf{X})) + J_{\lambda_n,\tau_n}(f)$ for any $f \in \mathcal{F}_n^{\mathrm{DNN}}$, we have

$$\begin{aligned} B_{2,n} &\le 2\left[\mathsf{P}_n\ell(Y\hat{f}_n(\mathbf{X})) - \mathsf{P}_n\ell(Yf_n^\circ(\mathbf{X}))\right] + 2J_{\lambda_n,\tau_n}(\hat{f}_n) \\ &\quad + 2\left[\mathsf{P}_n\ell(Yf_n^\circ(\mathbf{X})) - \mathsf{P}_n\ell(Yf^\star(\mathbf{X}))\right] \\ &\le 2J_{\lambda_n,\tau_n}(f_n^\circ) + 2\left[\mathsf{P}_n\ell(Yf_n^\circ(\mathbf{X})) - \mathsf{P}_n\ell(Y, f^\star(\mathbf{X}))\right] \end{aligned}$$

and so

$$\mathsf{E}(B_{2,n}) \le 2J_{\lambda_n,\tau_n}(f_n^\circ) + 2\mathcal{E}_{\mathsf{P}}(f_n^\circ) \le 2\inf_{f\in\mathcal{F}_n^{\mathrm{DNN}}}\left[\mathcal{E}_{\mathsf{P}}(f) + J_{\lambda_n,\tau_n}(f)\right] + \frac{1}{n}.$$

Combining all the bounds we have derived, we get the desired result. □

2.7.3 Proofs of Theorem 2.3.2 and Theorem 2.3.4

Proof of Theorem 2.3.2. Let $S_n = n^{\frac{1}{2\kappa+1}}\log^r n$. By Theorem 2.3.1, the assumption (2.3.6), and the fact that $\|\boldsymbol{\theta}(f)\|_{\mathsf{Clip},\tau} \le \|\boldsymbol{\theta}(f)\|_0$ for any $\tau > 0$, we have that for any $f^\star \in \mathcal{F}^\star$,

$$\begin{aligned} \mathsf{E}\left[\|\hat{f}_n - f^\star\|_{2,\mathsf{P}_\mathbf{X}}^2\right] &\lesssim \inf_{f\in\mathcal{F}_{0,n}^{\mathrm{DNN}}(S_n)}\left\{\|f - f^\star\|_{2,\mathsf{P}_\mathbf{X}}^2 + J_{\lambda_n,\tau_n}(f)\right\} + \frac{\log^2 n}{n} \\ &\lesssim \inf_{f\in\mathcal{F}_{0,n}^{\mathrm{DNN}}(S_n)}\|f - f^\star\|_{2,\mathsf{P}_\mathbf{X}}^2 + \lambda_n S_n + \frac{\log^2 n}{n} \\ &\lesssim (S_n/\log^r n)^{-2\kappa} + S_n\frac{\log^5 n}{n} + \frac{\log^2 n}{n} \\ &\lesssim n^{-\frac{2\kappa}{2\kappa+1}}\log^{5+r} n \end{aligned}$$

which concludes the desired result. □

Proof of Theorem 2.3.4. For $\mathbf{x} \in [0,1]^d$, define the function $\psi_\mathbf{x} : \mathbb{R} \mapsto \mathbb{R}_+$ by

$$\psi_\mathbf{x}(z) := \eta(\mathbf{x})\ell(z) - (1-\eta(\mathbf{x}))\ell(-z).$$

Note that $z_{\mathbf{x}}^\star := f_\ell^\star(\mathbf{x})$ is the minimizer of $\psi_{\mathbf{x}}(z)$ and satisfies $\psi_{\mathbf{x}}'(z_{\mathbf{x}}^\star) = 0$ $\mathsf{P}_{\mathbf{X}}$-a.s.. Then by the Taylor expansion around $z_{\mathbf{x}}^\star := f_\ell^\star(\mathbf{x})$

$$\psi_{\mathbf{x}}(z) - \psi_{\mathbf{x}}(z_{\mathbf{x}}^\star) = \frac{\psi''(\tilde{z})}{2}(z - z_{\mathbf{x}}^\star)^2$$

$\mathsf{P}_{\mathbf{X}}$-a.s., where $\tilde{z}$ lies between z and $z_{\mathbf{x}}^\star$. Since ℓ has a continuous second derivative, we have $\|\psi''\|_{\infty,[-F,F]} \leq c_{11}$ for some $c_{11} > 0$, which implies that

$$\mathcal{E}_{\mathsf{P}}(f) \leq c_{11}\|f - f_\ell^\star\|_{2,\mathsf{P}_{\mathbf{X}}}^2.$$

Let $S_n = n^{\frac{1}{2\kappa+1}} \log^r n$. By Theorem 2.3.3, the assumption (2.3.19), and the fact that $\|\boldsymbol{\theta}(f)\|_{\mathsf{Clip},\tau} \leq \|\boldsymbol{\theta}(f)\|_0$ for any $\tau > 0$, we have that for any $f_\ell^\star \in \mathcal{F}^\star$,

$$\begin{aligned}
\mathsf{E}\left[\mathcal{E}_{\mathsf{P}}(\hat{f}_n)\right] &\lesssim \inf_{f \in \mathcal{F}_{0,n}^{\mathrm{DNN}}(S_n)} \left\{\mathcal{E}_{\mathsf{P}}(f) + J_{\lambda_n,\tau_n}(f)\right\} + \frac{\log^2 n}{n} \\
&\lesssim \inf_{f \in \mathcal{F}_{0,n}^{\mathrm{DNN}}(S_n)} \left\|f - f_\ell^\star\right\|_{2,\mathsf{P}_{\mathbf{X}}}^2 + \lambda_n S_n + \frac{\log^2 n}{n} \\
&\lesssim n^{-\frac{2\kappa}{2\kappa+1}} \log^{3+r} n
\end{aligned}$$

which concludes the desired result. □

Chapter 3

Posterior consistency of the factor dimensionality in high-dimensional sparse factor models

3.1 Introduction

Factor models describe a dependence structure among correlated random variables in terms of a small number of unobserved variables called latent factors or just factors. To be specific, the (linear) factor model considered in this chapter assumes that a p-dimensional random vector $\mathbf{Y}$ is distributed as

$$\mathbf{Y}|\mathbf{Z} = \mathbf{z} \sim \mathsf{N}(\mathbf{Bz}, \sigma^2\mathbf{I}), \quad \mathbf{Z} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}) \tag{3.1.1}$$

where $\mathbf{B}$ is a $p \times K$ factor loading matrix, $\mathbf{Z}$ is a K-dimensional factor with $K < p$, and $\sigma^2 > 0$ is a noise variance. Under this model, the marginal distribution of $\mathbf{Y}$ is given by

$$\mathbf{Y} \sim \mathsf{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} := \mathbf{BB}^\top + \sigma^2\mathbf{I}.$$

That is, the distribution of $\mathbf{Y}$ is determined by the structured covariance matrix $\mathbf{B}\mathbf{B}^\top + \sigma^2\mathbf{I}$. This decomposition of the covariance matrix leads to the substantial reduction of the model complexity, and thus the factor model has been applied to a broad range of areas including high-dimensional covariance estimation [26, 30, 31], high-dimensional supervised learning [32, 53, 82, 86] and multiple testing under arbitrary dependence [27, 28, 57], and pupulary used in various application fields such as economy, psychology and gene expression studies [e.g., 9, 15, 33, 44, 65].

A major practical issue of using the factor model is to determine the *factor dimensionality* K. Frequentist approaches typically choose the factor dimensionality before estimating the loading matrix. One of the widely used methods is to fit the factor models for different values of K and to select the best K based on a model selection criterion [4, 5]. Alternatively, the factor dimensionality can be chosen based on the eigenvalues of the empirical covariance matrix [1, 55, 72].

Various priors have been developed for Bayesian analysis of the factor model with unknown factor dimensionality. Examples are spike and slab priors with the Indian buffet process (IBP) [17, 54, 79] and shrinkage type priors with the degree of shrinkage increasing across the column index [10, 83].

Large sample properties of the posterior distribution of the factor model also have received much attention. Pati et al. [75] investigated the posterior contraction rate of the covariance matrix with respect to the spectral norm for a sparse factor model where most of the entries in the factor loading matrix are zero. They showed that the derived posterior contraction rate is near-optimal in the minimax sense, up to a logarithm factor even when p is much larger than n. Xie et al. [99] obtained an improved contraction rate. However, they assumed that the true factor dimensionality is known to us.

Gao and Zhou [37] studied a Bayesian sparse principal component analysis (PCA) model, which is equivalent to the factor model with the constraint that the columns of the loading matrix are orthogonal to each other. They derived posterior contraction rates of the covariance matrix and the principal subspace estimation with respect to the spectral norm and proved the posterior consistency of the rank of the covariance matrix. Due to the orthogonality of the loading matrix, the rank of the covariance matrix is equal to the factor dimensionality. But there is no easy computational method to approximate the posterior distribution mainly because of the orthogonality

constraint on the loading matrix.

Rockova and George [79] considered the Bayesian factor model with a spike and slab prior with the IBP and proved that the posterior probability that the factor dimensionality is bounded by a certain quantity converges to one. But, the posterior consistency of the factor dimensionality is still unsolved.

In this chapter, we consider a Bayesian factor model with a spike and slab prior with the two-parameter IBP. We provide sufficient conditions on the prior for the posterior consistency of the factor dimensionality when the dimension p is very large, but the true factor loading matrix is sparse. Also, we derive the posterior contraction rate of the covariance matrix. Note that there is a straightforward but efficient Markov chain Monte carlo (MCMC) algorithm available for our Bayesian model. That is, our Bayesian model is the first one which not only is computationally tractable but also achieves the posterior consistency of the factor dimensionality.

This chapter is organized as follows. In Section 3.2, we introduce the assumptions on the true model. Also, we explain the proposed prior and its properties. In Section 3.3, we provide posterior asymptotic results. In Section 3.4, we conduct a simulation study to supplement our asymptotic results. In Section 3.5, we provide some discussions about an adaptiveness. Concluding remarks follow in Section 3.6. Finally, all the proofs are given in Section 3.7.

3.1.1 Notation

Let $\mathbb{R}$ be the set of real numbers and $\mathbb{N}$ be the set of natural numbers. For the positive integer p, we let $[p] := \{1, 2, \ldots, p\}$. For a real number x, $\lfloor x \rfloor$ denote the largest integer less than or equal to x and $\lceil x \rceil$ denote the smallest integer larger than or equal to x. For two positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ we write $a_n \lesssim b_n$ if there exists a positive constant $C > 0$ such that $a_n \leq C b_n$ for any $n \in \mathbb{N}$. Moreover, we write $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

For a set S, $|S|$ denotes its cardinality. For a p-dimensional vector $\boldsymbol{\beta} \equiv (\beta_j)_{j\in[p]}$, define $\boldsymbol{\beta}^S := (\beta_j : j \in S)$ for a subset $S \subset \{1, \ldots, p\}$. We let $\lambda_1(\boldsymbol{\Sigma}) \geq \lambda_2(\boldsymbol{\Sigma}) \cdots \geq \lambda_p(\boldsymbol{\Sigma})$ be the ordered eigenvalues of the $p \times p$ matrix $\boldsymbol{\Sigma}$. For a $p \times k$ matrix (k can be infinite) $\mathbf{A} = (a_{jh})_{j\in[p],h\in[k]}$, we denote the spectral

norm of the matrix $\mathbf{A}$ by $\|\mathbf{A}\|$ and the Frobenius norm by $\|\mathbf{A}\|_{\mathrm{F}}$. Recall that $\|\mathbf{A}\| := \sup_{\mathbf{x}\in\mathbb{R}^k:\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \lambda_1^{1/2}(\mathbf{A}^\top\mathbf{A})$ and $\|\mathbf{A}\|_{\mathrm{F}} := \sqrt{\mathsf{trace}(\mathbf{A}^\top\mathbf{A})}$. Also we use $\|\mathbf{A}\|_1$ to denote ℓ_1 norm of $\mathrm{vec}(\mathbf{A})$, i.e., $\|\mathbf{A}\|_1 := \sum_{j=1}^{p}\sum_{h=1}^{k}|a_{jh}|$. We let $\mathbf{A}_{>k} = (a_{jh})_{j\in[p],h\in\{k,k+1,\ldots,\}}$ and $\mathbf{A}_{\leq k} = (a_{jh})_{j\in[p],h\in[k]}$, which are the submatrices constructed by taking the colmuns with indices $> k$ and $\leq k$, respectively. Also we let $\mathbf{A}^S := (a_{jh})_{j\in S,h\in\mathbb{N}}$ for a subset $S \subset [p]$.

We denote by $\mathbb{1}(\cdot)$ the indicator function. Let $\Gamma(a)$ denote the gamma function and $\mathsf{B}(b,c)$ denote the beta function, where a, b and c are positive constants. Let $\mathbf{0}$ and $\mathbf{1}$ denote vectors of 0's and the one of 1's, respectively.

3.2 Assumptions and prior distribution

Throughout this chapter, we assume that for each $n \in \mathbb{N}$, we observe n independent and identically distributed p_n dimensional observations $\mathbf{Y}_{1:n} \equiv (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ and we model the data using the model (3.1.1).

3.2.1 Assumptions

In this section, we give the assumptions on the data generating process for asymptotic results of the posterior distribution.

Assumption 3.1. For each $n \in \mathbb{N}$, let $\mathbf{\Sigma}_{0n}$ be a true covariance matrix. We observe

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{\text{iid}}{\sim} \mathsf{N}_{p_n}(\mathbf{0}, \mathbf{\Sigma}_{0n})$$

where $\mathbf{\Sigma}_{0n}$ is of the form

$$\mathbf{\Sigma}_{0n} := \mathbf{B}_{0n}\mathbf{B}_{0n}^\top + \sigma_{0n}^2\mathbf{I}$$

where $\mathbf{B}_{0n} \in \mathbb{R}^{p_n\times k_{0n}}$. Here, k_{0n} is the true factor dimensionality.

We introduce some regularity conditions on the sequence of the true covariance matrices $\{\mathbf{\Sigma}_{0n}\}_{n\in\mathbb{N}}$.

Assumption 3.2. Assume that there exist sequences of positive real numbers $\{c_n\}_{n\in\mathbb{N}}$, $\{s_n\}_{n\in\mathbb{N}}$ satisfying the following conditions:

(A1) $\sum_{j=1}^{p_n} \mathbb{1}\left(\sum_{k=1}^{k_{0n}} |\beta_{0n,jk}| > 0\right) \leq s_n$, where $\beta_{0n,jk}$ denotes the (j,k)-th entry of $\mathbf{B}_{0n}$.

(A2) $\|\mathbf{\Sigma}_{0n}\| = c_n \lesssim s_n$.

(A3) $c_0 \leq \sigma_{0n}^2 \leq c_n$ for an universal constant $c_0 > 0$.

(A4) $c_n^2 s_n^2 k_{0n} \log p_n / n = o(1)$, $1 \leq k_{0n} < p_n/2$ and $p_n > n$.

(A5) $\lambda_{k_{0n}}\left(\mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right) > d_0$ for an universal constant $d_0 > 0$.

The assumption (A1) means that the number of nonzero rows of the true loading matrix is at most s_n. Since we consider a high dimensional case where p_n is much larger than n, we assume that the true loading matrix is sufficiently sparse to make the factor loading estimable.

The assumption (A2) implies that we allow the largest eigenvalue of $\mathbf{\Sigma}_{0n}$ to grow with the sample size. The bound $c_n \lesssim s_n$ is mild in view of random matrix theory. Suppose that $\tilde{B}$ be a $s_n \times k_{0n}$ random matrix whose entries are independent centered random variables with finite fourth moments. Then by Theorem 2 of [56], $\mathsf{E}\,\|\tilde{B}\| \lesssim \sqrt{s_n} + \sqrt{k_{0n}}$. If $k_{0n} \lesssim s_n$, we have that $\mathsf{E}\,\|\tilde{B}\tilde{B}^\top\| \lesssim s_n$. Pati et al. [75] and Rockova and George [79] assumed the same condition, while the other works on the Bayesian covariance estimation assumed that the largest eigenvalue of $\mathbf{\Sigma}_{0n}$ is bounded [37, 99]. If we assume the bounded largest eigenvalue, c_n disappears in the posterior convergence rate.

The assumption (A3) provides the lower and upper bound of the true noise variance σ_{0n}^2. The lower bound prevents that $\mathbf{\Sigma}_{0n}$ becomes ill-conditioned. Together with the assumption (A2) , the upper bound makes that the loading matrix $\mathbf{B}_{0n}$ dominates σ_{0n} in $\mathbf{\Sigma}_{0n}$.

The assumption (A4) gives a restriction on the quantities related to the true covariance matrices. We consider the high dimensional setups in which $p_n \gg n$ throughout. In particular, we can consider an ultra high dimensional case in which $p_n \asymp \exp(n^a)$ for some $a \in (0,1)$. We assume the quantity $c_n^2 s_n^2 k_{0n} \log p_n / n$ goes to zero as $n \to \infty$, which is equal to the square of the posterior contraction rate of the covariance matrix under our proposed

prior with respect to the spectral norm. In addition, we assume that the true model has at least one factor and the number of factors does not exceed half of the number of observed variables for technical reasons. (A1) and (A4) imply that the number of estimable parameters should grow slower than n.

The assumption (A5) creates and *eigengap* between the spikes and the noise, which prevents the underestimation of the factor dimensionality k_{0n}. The assumption (A5) plays a similar role as the *beta-min* condition which has been popularly used to prove variable section consistency in high-dimensional regression [16, 64].

3.2.2 Prior distribution and its properties

Let β_{jk} be the (j,k) entry of the $p \times \infty$-dimensional loading matrix $\mathbf{B}$. We consider a following prior distribution

$$\beta_{jk}|\xi_{jk} \overset{\text{ind}}{\sim} (1-\xi_{jk})\delta_0 + \xi_{jk}\mathsf{Laplace}(1),\ j \in [p],\ k \in \mathbb{N} \tag{3.2.1}$$

$$\xi_{jk}|\theta_k \overset{\text{ind}}{\sim} \mathsf{Bernoulli}\left(\theta_k\right),\ j \in [p],\ k \in \mathbb{N} \tag{3.2.2}$$

$$\theta_k := \prod_{h=1}^{k} \nu_h,\ k \in \mathbb{N} \text{ where } \nu_h \overset{\text{iid}}{\sim} \mathsf{Beta}(\alpha, \kappa+1),\ h \in \mathbb{N}. \tag{3.2.3}$$

where $\alpha > 0$ and $\kappa \geq 0$ are hyperparameters. Note that we use the stick-breaking representation of the two-parameter IBP [90] here. We refer to the above distribution on $\mathbf{B}$ as $\mathsf{SSIBP}_p(\alpha,\kappa)$, which is an abbreviation of *spike and slab Indian buffet process*. We denote by $\Pi(\cdot)$ the prior distribution of $\mathsf{SSIBP}_p(\alpha,\kappa)$.

The (one-parameter) IBP introduced by [38], which is the prior on ξ_{jk} with $\kappa = 0$ has been popularly used as a prior on the nonzero entries of the loading matrix. The prior distribution of [54] for $\mathbf{B}$ is almost the same as SSIBP except that κ is set to be 0 and the Laplace distribution in (3.2.1) is replaced by the normal distribution. Rockova and George [79] used the IBP for the prior of ξ_{jk}, but they replace δ_0 in (3.2.1) with the Laplace distribution with a very small dispersion. This replacement enables us to use the fast and scalable expectation-maximization algorithm that estimates a posterior mode.

Note that both Knowles and Ghahramani [54] and Rockova and George [79] considered the one-parameter IBP which is equivalent to the two-parameter IBP with $\kappa = 0$. We introduce the additional hyperparameter κ in order to get the posterior consistency of the factor dimensionality by choosing κ appropriately. Although some works [17, 73] used the two-parameter IBP, no theoretical study has been done.

In the following three subsections, we present lemmas which provide some theoretical properties of our considered prior.

3.2.2.1 Induced distribution of the factor dimensionality

In this section, we derive an upper bound of the tail probability of the factor dimensionality induced by the SSIBP prior. We first define the factor dimensionality of the $p \times \infty$ dimensional loading matrix $\mathbf{B}$.

Definition 3.2.1. For a given $p \times \infty$ loading matrix $\mathbf{B} \equiv (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots)$, we define the *factor dimensionality* $\mathsf{K}^+(\mathbf{B})$ as the number of nonzero columns of $\mathbf{B}$, i.e.,

$$\mathsf{K}^+(\mathbf{B}) := \sum_{k=1}^{\infty} \mathbb{1}\left(\|\boldsymbol{\beta}_k\|_0 \geq 1\right)$$

where $\boldsymbol{\beta}_k$ denotes the k-th column of $\mathbf{B}$.

The following lemma shows that the tail probability of the factor dimensionality is exponentially decaying as the factor dimensionality increases.

Lemma 3.2.1. *If* $\mathbf{B} \sim \mathsf{SSIBP}_p(\alpha, \kappa)$ *for* $\alpha >$ *and* $\kappa \geq 0$, *then for any* $k \in \mathbb{N}$

$$\Pi(\mathsf{K}^+(\mathbf{B}) > k) \leq C_{\alpha,\kappa} p \left(\frac{\alpha}{\alpha + \kappa + 1}\right)^{k+1}, \tag{3.2.4}$$

where $C_{\alpha,\kappa} := 2\left((\alpha + \kappa + 1)/(\kappa + 1) + 4/3\right)$. *In particular, if* $\alpha \leq \kappa + 1$, *then*

$$\Pi(\mathsf{K}^+(\mathbf{B}) > k) \leq 6p \left(\frac{\alpha}{\alpha + \kappa + 1}\right)^{k+1}. \tag{3.2.5}$$

Proof. The proof is deferred to Section 3.7.1. □

3.2.2.2 Induced distribution of the sparsity

We use a notion of row-wise sparsity of the loading matrix. This notion of sparsity is also used to define the sparsity of the true loading matrix in the assumption (A1).

Definition 3.2.2. For a $p \times \infty$ loading matrix $\mathbf{B} \equiv (\beta_{jh})_{j\in[p],h\in\mathbb{N}}$ and a positive integer $k \in \mathbb{N}$, we define the *row-support up to k-th column* of $\mathbf{B}$ as

$$\mathsf{supp}_k(\mathbf{B}) := \left\{ j \in [p] : \sum_{h=1}^{k} |\beta_{jh}| > 0 \right\}.$$

$\mathbf{B}$ is said to be *s-sparse up to k* if $|\mathsf{supp}_k(\mathbf{B})| \leq s$.

The row-wise sparsity is considered by [37] and [99], While the loading matrices in [37] and [99] have finite columns, we work with loading matrices with infinite columns and thus need a truncated version of the row-wise sparsity.

Throughout this chapter, we set $\kappa = p^{1+\delta}$ for a fixed constant $\delta > 0$. This choice of the prior parameter is intended to put most of the prior mass concentrating on sufficiently (row-wise) sparse loading matrices, which is shown in the following lemma.

Lemma 3.2.2. *If* $\mathbf{B} \sim \mathsf{SSIBP}_p(\alpha, p^{1+\delta})$ *with* $\alpha \in (0,1)$ *and* $\delta > 0$, *then for any* $k \in \mathbb{N}$, *and* $t \geq 1$

$$\Pi(|\mathsf{supp}_k(\mathbf{B})| > t) \leq (k+1)\mathrm{e}^{-(\delta/12)t\log p}. \tag{3.2.6}$$

Proof. The proof is deferred to Section 3.7.1. □

3.2.2.3 Prior concentration near the true loading matrix

In this subsection, we show that the proposed prior puts sufficiently large mass near the truth. Since the true loading matrix depends on n, we let also prior parameters α and κ also depends on the sample size n. We let $\Pi_n(\cdot)$ be the prior distribution of $\mathsf{SSIBP}_{p_n}(\alpha_n, \kappa_n)$. Unless there is a confusion, we understand the loading matrix $\mathbf{B}_{0n} \in \mathbb{R}^{p_n \times k_{0n}}$ as the $p_n \times \infty$ dimensional

matrix $(\mathbf{B}_{0n}, \mathbf{0}_{p_n \times \infty})$, where $\mathbf{0}_{p_n \times \infty}$ denotes $p_n \times \infty$ dimensional matrix of zeros.

Lemma 3.2.3. *Suppose that $\mathbf{B}_{0n} \in \mathbb{R}^{p_n \times k_{0n}}$ be a s_n-sparse up to k_{0n} loading matrix. Let $\mathbf{B} \sim \mathsf{SSIBP}_{p_n}(\alpha_n, p_n^{1+\delta})$ for $\delta > 0$. Then for any $n \in \mathbb{N}$ and $\eta_n > 0$,*

$$\begin{aligned} &\Pi_n\left(\|\mathbf{B} - \mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n\right) \\ &\quad \geq \alpha_n^{k_{0n}} \exp\left(-\|\mathbf{B}_{0n}\|_1 - \eta_N k_{0n} - C_1 s_n k_{0n} \log\left(p_n \vee \eta_n^{-1}\right)\right), \end{aligned} \tag{3.2.7}$$

for some universal constant $C_1 > 0$ depending only on δ. Moreover, if the assumptions (A2) and (A3) hold and $\eta_n \lesssim 1$, we have

$$\Pi_n\left(\|\mathbf{B} - \mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n\right) \geq \alpha_n^{k_{0n}} \exp\left(-C_2 s_n k_{0n} \log\left(p_n \vee \eta_n^{-1}\right)\right), \tag{3.2.8}$$

for some universal constant $C_2 > 0$ depending only on δ.

Proof. The proof is deferred to Section 3.7.1. □

Most priors for high dimensional sparse factor models have lower bound

$$\exp(-C_3 s_n k_{0n} \log\left(p_n \vee \eta_n^{-1}\right))$$

for some universal constant $C_3 > 0$ to the prior concentration [37, 75, 79, 99]. The lower bound in (3.2.7) is similar to them, but key difference of (3.2.7) is that the lower bound depends explicitly on the prior parameter α_n. For the posterior consistency of factor dimensionality, controlling α_n is indispensable.

Using Lemma 3.2.3, we can obtain the following prior concentration result for the sequence of true covariance matrices with respect to the Frobenius norm.

Corollary 3.2.4. *Suppose that $\mathbf{\Sigma}_{0n}$ satisfies (A1)-(A4). Let $\mathbf{B} \sim \mathsf{SSIBP}_{p_n}(\alpha_n, p_n^{1+\delta})$ and $\sigma^2 \sim \mathsf{IG}(a, b)$ for $\delta > 0$, $a > 0$ and $b > 0$. Then,*

$$\Pi_n\left(\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\|_{\mathrm{F}} \leq \sqrt{\frac{s_n k_{0n}}{n}}\right) \geq \alpha_n^{k_{0n}} \mathrm{e}^{-C_1 s_n k_{0n} \log p_n}, \tag{3.2.9}$$

for some universal constant $C_1 > 0$ depending only on δ, a and b.

Proof. The proof is deferred to Section 3.7.1. □

3.3 Asymptotic properties of the posterior distribution

In this section we study asymptotic properties of the posterior distributions of the covariance matrix and the factor dimensionality in the sparse factor model, respectively.

Given data $\mathbf{Y}_{1:n} \equiv (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, we denote by $\Pi_n(\cdot|\mathbf{Y}_{1:n})$ the posterior distribution induced by the prior Π_n and the data $\mathbf{Y}_{1:n}$. Let $\sigma(\mathbf{Y}_{1:n})$ be the σ-field generated by $\mathbf{Y}_{1:n}$. For a given sample size n and covariance matrix $\mathbf{\Sigma}$, we let $\mathsf{P}_{\mathbf{\Sigma}}$ and $\mathsf{E}_{\mathbf{\Sigma}}$ denote the probability measure and the expectation operator under the law $(\mathsf{N}(\mathbf{0}, \mathbf{\Sigma}))^n$, where we suppress the dependence on n for simplicity.

3.3.1 Posterior contraction rate for covariance matrix

We let $\mathcal{C}_{0n}$ be a sequence of the classes of covariance matrices such that

$$\mathcal{C}_{0n} := \left\{ \mathbf{\Sigma}_{0n} \equiv \mathbf{B}_{0n}\mathbf{B}_{0n}^\top + \sigma_{0n}^2\mathbf{I} : \text{ (A1)- (A4) are satisfied} \right\}.$$

Note that we do not require the assumption (A5) in the definition of $\mathcal{C}_{0n}$. The next theorem derives the posterior contraction rate of the covariance matrix with respect to the spectral norm.

Theorem 3.3.1. *A prior, let $\mathbf{B} \sim \mathsf{SSIBP}_{p_n}(\alpha_n, p_n^{1+\delta})$ with $\alpha_n < 1$ for any sufficiently large n and that $\sigma^2 \sim \mathsf{IG}(a, b)$ for $\delta > 0$, $a > 0$ and $b > 0$. Then for any sufficiently large $M > 0$,*

$$\sup_{\mathbf{\Sigma}_{0n} \in \mathcal{C}_{0n}} \mathsf{E}_{\mathbf{\Sigma}_{0n}} \left[\Pi_n \left(\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| > M\epsilon_n \middle| \mathbf{Y}_{1:n} \right) \right] = o(1), \tag{3.3.1}$$

where

$$\epsilon_n := c_n \sqrt{\frac{k_{0n}}{n} \max\left\{ s_n \log p_n, \log\left(\frac{1}{\alpha_n}\right) \right\}}. \tag{3.3.2}$$

Proof. The proof is deferred to Section 3.7.2. □

If we set the hyperparameter α_n to satisfy $\log\left(\frac{1}{\alpha_n}\right) \lesssim s_n \log p_n$, the posterior contraction rate becomes

$$\epsilon_n = c_n \sqrt{\frac{s_n k_{0n} \log p_n}{n}} \tag{3.3.3}$$

which is minimax optimal [75]. For example, when $\alpha_n = \alpha_0$ for some $\alpha_0 \in (0, 1)$ or $\alpha_n = p_n^{-1}$ as in [79], the posterior contraction rates becomes minimax optimal.

Pati et al. [75] derived a similar posterior contraction rate with their own prior and assumptions. In our terminology, the posterior contraction rate of [75] is equivalent to $\sqrt{\log n} c_n \sqrt{s_n k_{0n}^2 \log p_n / n}$. We remove $\sqrt{k_{0n} \log n}$ factor by using the improved test construction used in [37, 99].

Gao and Zhou [37] obtained the posterior contraction rate $\sqrt{s_n k_{0n} \log p_n / n}$ under the assumption that the largest eigenvalue of the true covariance matrix is bounded and Xie et al. [99] obtained the posterior contraction rate of $\sqrt{s_n \log p_n / n}$ under the assumptions that both the true largest eigenvalue and the true factor dimensionality are bounded. Our posterior contraction rate (3.3.3) recovers those rates under the additional assumptions they made.

3.3.2 Posterior consistency of the factor dimensionality

We let $\mathcal{C}_{0n}^\star$ be a sequence of the classes of covariance matrices such that

$$\mathcal{C}_{0n}^\star := \left\{ \mathbf{\Sigma}_{0n} \equiv \mathbf{B}_{0n} \mathbf{B}_{0n}^\top + \sigma_{0n}^2 \mathbf{I} : \text{ (A1)- (A5) are satisfied} \right\},$$

which is a subset of $\mathcal{C}_{0n}$. We need an additional condition (A5) for the posterior consistency of the factor dimensionality.

The following theorem proves that the posterior consistency of the factor dimensionality.

Theorem 3.3.2. *A priori, let* $\mathbf{B} \sim \mathsf{SSIBP}_{p_n}(p_n^{-As_n^2}, p_n^{1+\delta})$ *for sufficiently large* $A > 0$ *and that* $\sigma^2 \sim \mathsf{IG}(a, b)$ *for* $\delta > 0$, $a > 0$ *and* $b > 0$. *Then*

$$\sup_{\mathbf{\Sigma}_{0n} \in \mathcal{C}_{0n}^{\star}} \mathsf{E}_{\mathbf{\Sigma}_{0n}} \left[\Pi_n \left(\mathsf{K}^{+}(\mathbf{B}) \neq k_{0n} \middle| \mathbf{Y}_{1:n} \right) \right] = o(1). \tag{3.3.4}$$

Proof. The proof is deferred to Section 3.7.2. □

Rockova and George [79] proved that

$$\mathsf{E}_{\mathbf{\Sigma}_{0n}} \left[\Pi_n \left(\mathsf{K}^{+}(\mathbf{B}) > M s_n k_{0n} \middle| \mathbf{Y} \right) \right] = o(1)$$

for our prior distribution with $\kappa_n = 0$ (i.e., the one-parameter IBP). This result is weaker than ours when s_n diverges. Also Rockova and George [79] did not consider the posterior probabilities of the underestimation of the factor dimensionality. We introduce the assumption (A5) and use a diverging value of κ_n to get the posterior consistency of the factor dimensionality.

For the α_n in Theorem 3.3.2, the posterior contraction rate of the covariance matrix becomes

$$\epsilon_n = c_n \sqrt{\frac{s_n^2 k_{0n} \log p_n}{n}}. \tag{3.3.5}$$

which is s_n times slower than the optimal rate.

3.4 Numerical results

To illustrate our theoretical foundings, we compare the concentrations of the posterior distributions for the factor dimensionality and the covariance matrix with various choices of the hyperparameters α_n and κ_n by simulation.

3.4.1 MCMC algorithm

Let $K^* := \mathsf{K}^+(\mathbf{B})$ be the number of nonzero columns of the loading matrix.

For $j \in [p]$ and $k \in [K^*]$, the factor loading β_{jk} is sampled from the conditional posterior

$$\beta_{jk}|- \sim \begin{cases} \mathsf{N}(\hat{\beta}_{jk}, \hat{\tau}_{jk}) & \text{if } \xi_{jk} = 1 \\ \delta_0 & \text{if } \xi_{jk} = 0 \end{cases}$$

where

$$\widehat{\tau}_{jk} := \left(\sigma^{-2} \sum_{i=1}^{n} Z_{ik}^2 + \tau_{jk}^{-1} \right)^{-1}$$
$$\widehat{\beta}_{jk} := \widehat{\tau}_{jk} \left\{ \sigma^{-2} \sum_{i=1}^{n} Z_{ik} \left(Y_{ij} - \sum_{h \in [K^*]: h \neq k} Z_{ih} \beta_{jh} \right) \right\}.$$

For $j \in [p]$ and $k \in [K^*]$, the auxiliary parameter τ_{jk} is sampled from

$$\tau_{jk}|- \sim \begin{cases} \mathsf{GIG}(1, \beta_{jk}^2, \frac{1}{2}) & \text{if } \xi_{jk} = 1 \\ \mathsf{Exp}(\frac{1}{2}) & \text{if } \xi_{jk} = 0 \end{cases}$$

where $\mathsf{GIG}(a, b, p)$ denotes the generalized inverse Gaussian (GIG) distribution with the density $f(z)$ proportional to $f(z) \propto z^{p-1} \mathrm{e}^{-(az+b/z)/2}$.

For $j \in [p]$, the indicator parameters are sampled as follows. For $k \in [K^*]$, ξ_{jk} is sampled with probability

$$\frac{\Pi(\xi_{jk} = 1|-)}{\Pi(\xi_{jk} = 1|-)} = \frac{\alpha + p_{jk}}{\kappa + p - p_{jk}} \sqrt{\frac{\widehat{\tau}_{jk}}{\tau_{jk}}} \exp\left(\frac{1}{2\widehat{\tau}_{jk}} \widehat{\beta}_{jk}^2 \right).$$

where $p_{jk} := \sum_{l\in[p]:l\neq j} \xi_{lk}$. For $k > K^*$, we first propose $K_j^* \in \mathbb{N}\cup\{0\}$ and $\beta_j^* \in \mathbb{R}^{K_j^*}$ from the proposal distribution

$$J(K_j^*)J(\beta_j^*|K_j^*) = \mathsf{Poisson}(1)\mathsf{Lap}(1)^{K_j^*}.$$

Then accept the proposal with probability

$$\max\left\{1, |2\pi\mathbf{M}_j|^{-n/2}\, \mathrm{e}^{u_j^2(\beta_j^*)^\top \mathbf{M}_j^{-1}\beta_j^*/2} \frac{\mathrm{e}^{1-\frac{\alpha}{p-1}}\left(\frac{\alpha}{p-1}\right)^{K_j^*}}{K_j^*!}\right\} \tag{3.4.1}$$

where

$$\begin{aligned}\mathbf{M}_j &:= \sigma^{-2}\beta_j^*(\beta_j^*)^\top + \mathbf{I}\\ u_j &:= \sigma^{-2}\sum_{i=1}^{n}\left(Y_{ij} - \sum_{k=1}^{K^*} Z_{ik}\beta_{jk}\right).\end{aligned}$$

If the proposal is accepted, we update

$$\begin{aligned}\mathbf{B} &\leftarrow (\mathbf{B}, (\beta_{jk}^*\mathbb{1}(l=j))_{l\in[p],k\in[K_j^*]})\\ K^* &\leftarrow K^* + K_j^*\end{aligned}$$

For $i \in [n]$, the latent variable $\mathbf{Z}_i$ is sampled from

$$\mathbf{Z}_i|- \sim \mathsf{N}\left(\sigma^{-2}\widehat{\mathbf{\Sigma}}_{\mathbf{Z}}\mathbf{B}^\top\mathbf{Y}_i, \widehat{\mathbf{\Sigma}}_{\mathbf{Z}}\right)$$

where

$$\widehat{\mathbf{\Sigma}}_{\mathbf{Z}} := (\sigma^{-2}\mathbf{B}^\top\mathbf{B} + \mathbf{I})^{-1}.$$

The noise variance parameter σ^2 is sampled from

$$\sigma^2|- \sim \mathsf{IG}\left(a + \frac{np}{2}, b + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(Y_{ij} - \sum_{k=1}^{K^*} Z_{ik}\beta_{jk}\right)^2\right).$$

3.4.2 Simulation study

We let the sample size n vary among $\{50, 150, 250\}$. For each sample size n, we let the dimension of the data be equal to $p_n = 2n$ and the number of nonzero rows be equal to $s_n = \left\lfloor \sqrt{p_n} \right\rfloor$. We consider the fixed factor dimensionality cases where the true factor dimensionality k_{0n} varies among $\{1, 5\}$. For given s_n and k_{0n}, the loading matrix $\mathbf{B}_{0n}$ is generated by first selecting s nonzero rows randomly and sampling loadings in the first k_0 columns and nonzero rows from $\{-2, 0, 2\}$ randomly. A noise variance is set to be 1. The simulated data are generated by the Gaussian distribution with the generated covariance matrix $\mathbf{\Sigma}_{0n} := \mathbf{B}_{0n}\mathbf{B}_{0n}^\top + \mathbf{I}$.

We consider two different choices of α_n. The first one is $\alpha_n = p_n^{-1}$, which is used by [79]. The other one is $\alpha_n = p_n^{-s_n^2}$ whose use is advocated by our theory. In addition, we consider two different choices of κ_n. The first one is $\kappa_n = 0$, which corresponds to the one-parameter IBP. The other one is $\kappa_n = p_n^{1+\delta}$ with $\delta = 0$, which we recommend to use.

We compare the posterior probability of correct estimation of the factor dimensionality, i.e., $\Pi(\mathsf{K}^+(\mathbf{B}) = k_{0n} | \mathbf{Y}_{1:n})$, and the scaled spectral norm loss $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{0n}\| / \|\mathbf{\Sigma}_{0n}\|$, where $\hat{\mathbf{\Sigma}}$ is the posterior mean of the covariance matrix. We repeat the simulation 50 times, and report averages of those quantities across simulation replicates.

Figure 3.1 presents the averaged fraction of correct estimation of the factor dimensionality for the tour choices of the hyperparameters. We can recover the factor dimensionality more precisely with the extreme choice $\alpha_n = p_n^{-s_n^2}$. The mild choice $\alpha_n = p_n^{-1}$ leads to the posterior distribution that puts most of its mass to the larger values than the true factor dimensionality. For the hyperparameter κ_n, the value $\kappa_n = p_n^{1+\delta}$ yields higher posterior concentration than the value $\kappa_n = 0$.

Figure 3.2 presents the averaged scaled spectral norm loss for covariance matrix estimation. The results are almost similar for all the choices of the hyperparameters. It supports the fact that when we use $\alpha_n = p_n^{-s_n^2}$, the posterior contraction rate worsens by only $\sqrt{s_n}$ factor, which is not critical compared to the overall contraction rate.

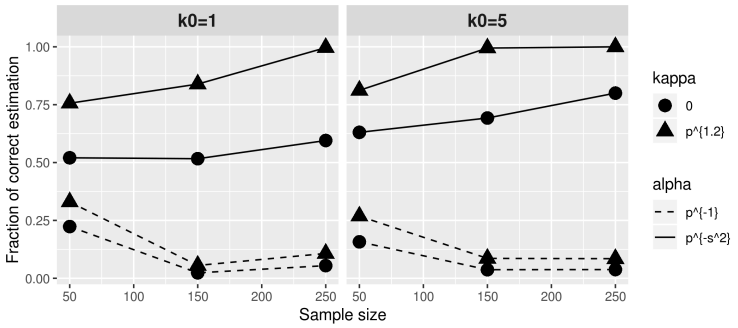


FIGURE 3.1: Fraction of correct estimation of the factor dimensionality by the value of the hyperparemters α_n and κ_n. The average value across simulation replications are plotted versus the sample size for $k_0 = 1$ (*left*), and $k_0 = 5$ (*right*).

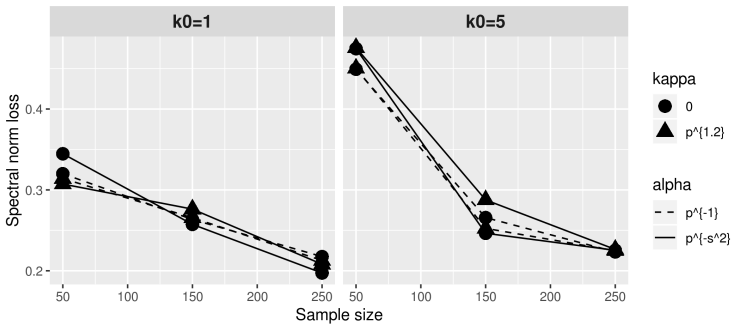


FIGURE 3.2: The scaled spectral norm loss for the covariance matrix estimation by the value of the hyperparemters α_n and κ_n. The average value across simulation replications are plotted versus the sample size for $k_0 = 1$ (*left*), and $k_0 = 5$ (*right*).

3.5 Discussions about adaptive priors

For the posterior consistency of the factor dimensionality with the SSIBP prior, we need to know the true sparsity level s_n when we set the hyperparameters. However in general s_n is not known in practice. It would be useful to have the posterior consistency without knowing the true sparsity level s_n, In this section, we briefly discuss how to construct such adaptive priors to attain the posterior consistency of the factor dimensionality.

The sparse PCA prior considered by Gao and Zhou [37] is an adaptvie prior. Gao and Zhou [37] proves that their spar PCA prior is adaptive. However, their prior imposes the orthogonality of the column vectors of the loading matrix. That is, the sparse PCA prior satisfies

$$\Pi\left(\bigcap_{k=1}^{k^*}\left\{\left\|\boldsymbol{\beta}_k\right\|_2 \geq d^* \text{ and } \boldsymbol{\beta}_h^\top \boldsymbol{\beta}_k = 0, \forall h \neq k\right\}\right) = 1 \tag{3.5.1}$$

for some $d^* > 0$, where k^* is the upper bound to the number of columns of the loading matrices. This orthogonality of the column vectors makes the posterior contration of the covariance matrix automaticaly imply the posterior consistiency of the factor dimensionality, and for the posterior contraction rate of the covariance matrix, we do not need to know s_n. But the complicated nature of condition (3.5.1) makes it difficult to construct an efficient MCMC sampler.

As an alternative we consider a two-step prior, denoted by $\check{\Pi}_n$, on the loading matrix $\mathbf{B} \equiv (\beta_{jk})_{j\in[p_n],k\in\mathbb{N}}$. The two-step prior first selects the number of nonzero columns K from $\mathbb{N}$ and a size s of the row-support up to column K from $[p]$ according to a certain distribution $\check{\Pi}_{n,1}$. and selects a random subset $S \subset [p]$ of cardinality $|S| = s$ with equal probability. Then $\{\beta_{jh} : j \in S, h \in [k]\} \sim \check{\pi}_{n,2}$ for a certain distribution $\check{\pi}_{n,2}$ on $\mathbb{R}^{s\times k}$ and sets $\beta_{jh} = 0$ if $j \notin S$ or $h \geq k+1$. This two-step prior can be written as

$$(\mathbf{B}, S, k) \mapsto \check{\Pi}_{n,1}(|S|, k)\frac{1}{\binom{p}{|S|}}\check{\pi}_{n,2}\left(\mathbf{B}^S_{\leq K}\right)\delta_{\mathbf{0}}\left(\mathbf{B}^{S^c}_{\leq K}\right)\delta_{\mathbf{0}}\left(\mathbf{B}_{>K}\right), \tag{3.5.2}$$

where $\delta_{\mathbf{0}}$ denotes the point mass at 0, $\mathbf{B}^S_{\leq K} := (\beta_{jk})_{j\in S,k\in[K]}$ and $\mathbf{B}^{S^c}_{\leq K}$ and $\mathbf{B}_{>K}$ are defined similarly.

The suitable condition on $\check{\Pi}_{n,1}$ for the adaptiveness of the posterior consistency of the factor dimensionality, analog to the condition (P1) in [42], is

$$\check{\Pi}_{n,1}(|S| = s^*, K = k^*) \propto \exp(-A(s^*)^2 k^* \log p) \tag{3.5.3}$$

for some $A > 0$. Under this condition, we can prove the posterior consistency for the factor dimensionality in Theorem 3.5.1.

Theorem 3.5.1. *Let $b_0 > 0$ be the fixed constant. Define the class $\mathcal{C}_{0n}^{\star\star}$ of covariance matrices by*

$$\mathcal{C}_{0n}^{\star\star} := \left\{ \mathbf{\Sigma}_{0n} \equiv \mathbf{B}_{0n}\mathbf{B}_{0n}^\top + \sigma_{0n}^2 \mathbf{I} \in \mathcal{C}_{0n}^{\star} : \min_{\beta_{0n,jk}:\beta_{0n,jk}\neq 0} \left|\beta_{0n,jk}\right| > b_0 \right\}.$$

Assume that the loading matrix $\mathbf{B} \equiv (\beta_{jk})_{j\in[p_n],k\in\mathbb{N}}$ follows the prior $\check{\Pi}_n$ of (3.5.2) and $\sigma^2 \sim \mathsf{IG}(a,b)$ for $a > 0$ and $b > 0$. Suppose that $\check{\Pi}_{n,1}$ is the distribution satisfying that for any $(s^*, k^*) \in [p] \times \mathbb{N}$,

$$\check{\Pi}_{n,1}(|S| = s^*, K = k^*) \propto \mathrm{e}^{-A(s^*)^2 k^* \log p_n} \tag{3.5.4}$$

for sufficiently large $A > 0$, and $\check{\pi}_{n,2}$ is the product of densities such that

$$\check{\pi}_{n,2}\left(\mathbf{B}_{\leq K}^S\right) = \prod_{(j,k)\in S\times[K]} g(\beta_{jk}) \tag{3.5.5}$$

where $g(\cdot)$ denotes the density of $\mathsf{Laplace}(1)$. Then

$$\sup_{\mathbf{\Sigma}_{0n}\in\mathcal{C}_{0n}^{\star\star}} \mathsf{E}_{\mathbf{\Sigma}_{0n}}\left[\check{\Pi}_n\left(\mathsf{K}^+(\mathbf{B}) \neq k_{0n} \middle| \mathbf{Y}_{1:n}\right)\right] = o(1),$$

$$\sup_{\mathbf{\Sigma}_{0n}\in\mathcal{C}_{0n}^{\star\star}} \mathsf{E}_{\mathbf{\Sigma}_{0n}}\left[\check{\Pi}_n\left(\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| > Mc_n\sqrt{\frac{s_n^2 k_{0n}\log p_n}{n}} \middle| \mathbf{Y}_{1:n}\right)\right] = o(1),$$

for sufficiently large $M > 0$, where $\check{\Pi}_n(\cdot|\mathbf{Y}_{1:n})$ denotes the posterior distribution induced by the prior $\check{\Pi}$.

Proof. The proof is deferred to Section 3.7.3. □

Remark 3.5.1. For the SSIBP prior, it can be shown that

$$\begin{aligned}\Pi_n\left(\left|\mathsf{supp}_{k^*}(\mathbf{B})\right| = s^*, \mathsf{K}^+(\mathbf{B}) = k^*\right) &\propto \alpha^{k^*} \exp(-Cs^* \log p_n) \\ &= \exp(-Cs^* \log p_n - k^* \log(1/\alpha))\end{aligned}$$

and hence the condition (3.5.3) is not satisfied.

Remark 3.5.2. There is the additional assumption $\min_{\beta_{0n,jk}:\beta_{0n,jk}\neq 0} |\beta_{0n,jk}| > b_0$ in $\mathcal{C}_{0n}^{\star\star}$ of Theorem 3.5.1. The constant b_0 can be replaced by a positive sequence $(b_{0n})_{n\in\mathbb{N}}$ going to 0 with a slower rate than the posterior contraction rate of the covariance matrix.

Even though the two-step prior considered in Theorem 3.5.1 has good theoretical properties, it would not be straightforward to construct a computationally tractable posterior inference algorithm. Reversible jump MCMC could be used for this purpose. We leave this problem as a future work.

3.6 Concluding remarks

In this chapter, we proposed a computationally tractable prior which has desirable large sample properties. The proposed prior enables consistent estimation of both the factor dimensionality and induced covariance matrix. We also derived that the posterior contraction rate for covariance matrices, which is slightly slower than the minimax convergence rate by a multiple of the true sparsity level.

The proposed prior distribution is nonadaptive supposing that the true sparsity level is known to us. Gao and Zhou [37] proved that their sparse PCA prior adaptively attains the posterior consistency of the factor dimensionality. But it is not evident that there is a computationally tractable posterior inference algorithm. We are plan to develop a both computationally tractable and adaptive prior in a near future.

In an ultra high dimensional setup in which the dimension p_n grows exponentially in n, the number of nonzero rows s_n of the loading matrix

should be of order $O(\log p_n)$. This order is quite small compared to the dimension p_n, which means that the sparse factor model is not suitable for high dimensional correlated data. A G-block covariance model considered by [12] can be an alternative, but there is no Bayesian estimation methods with theoretical guarantees. We investigate this problem as a future work.

3.7 Proofs

3.7.1 Proofs of lemmas and corollary in Section 3.2

Proof of Lemma 3.2.1. If every β_{jh} for $j \in [p]$ and $h \geq k+1$ is equal to 0, it holds that $\mathsf{K}^+(\mathbf{B}) \leq k$. Hence, by Lemma 3.7.1,

$$\begin{aligned}\mathsf{P}(\mathsf{K}^+(\mathbf{B}) \leq k) &\geq \mathsf{E}\left(\prod_{h=k+1}^{\infty}(1-\theta_h)^p\right) \\ &\geq \exp\left\{-2\left(p\frac{\alpha+\kappa+1}{\kappa+1}+\frac{4}{3}\right)\left(\frac{\alpha}{\alpha+\kappa+1}\right)^{k+1}\right\}.\end{aligned}$$

Since $2\left(p\frac{\alpha+\kappa+1}{\kappa+1}+\frac{4}{3}\right) \leq 2p\left(\frac{\alpha+\kappa+1}{\kappa+1}+\frac{4}{3}\right)$, we have

$$\begin{aligned}\mathsf{P}(\mathsf{K}^+(\mathbf{B}) > k) &\leq 1-\exp\left(-C_{\alpha,\kappa}p\left(\frac{\alpha}{\alpha+\kappa+1}\right)^{k+1}\right) \\ &\leq C_{\alpha,\kappa}p\left(\frac{\alpha}{\alpha+\kappa+1}\right)^{k+1},\end{aligned}$$

which completes the proof. □

Proof of Lemma 3.2.2. Since the inequality follows trivially when $t > p$, we assume $t \leq p$. Let $(\theta_h)_{h\in[k]}$ be given. Then the random variable $S_k :=$

$|\mathsf{supp}_k(\mathbf{B})|$ is distributed as $\mathsf{Binom}(p, \pi_\theta)$ with

$$\begin{aligned}\pi_\theta &:= \Pi\left(\sum_{h=1}^{k} |\beta_{1h}| > 0 \middle| (\theta_h)_{h\in[k]}\right) \\ &= 1 - \Pi\left(\beta_{1h} = 0, \forall h \in [k] | (\theta_h)_{h\in[k]}\right) \\ &= 1 - \prod_{h=1}^{k}(1-\theta_h) \leq \sum_{h=1}^{k}\theta_h.\end{aligned}$$

Recall that $\theta_h = \prod_{l=1}^{h} \nu_l$ with $\nu_l \overset{\text{iid}}{\sim} \mathsf{Beta}(\alpha, \kappa+1)$. Define the set

$$\mathcal{E}_k := \left\{(v_l)_{l\in\mathbb{N}} : v_l \leq \frac{t\delta \log p}{7p^{1+\delta}}, \forall l \in [k]\right\}.$$

We bound $\Pi(S_k \geq t)$ as

$$\Pi(S_k \geq t) \leq \mathsf{E}\left[\Pi\left(S_k \geq t | (\theta_h)_{h\in[k]}\right) \mathbb{1}_{\mathcal{E}_k}\right] + \Pi(\mathcal{E}_k^c).$$

On the event $\mathcal{E}_k$, we have that

$$\begin{aligned}\pi_\theta \leq \sum_{h=1}^{k}\theta_h &\leq \sum_{h=1}^{k}\left(\frac{t\delta\log p}{7p^{1+\delta}}\right)^h \\ &\leq \sum_{h=1}^{\infty}\left(\frac{t\delta\log p}{7p^{1+\delta}}\right)^h \\ &\leq \frac{t\delta\log p}{7p^{1+\delta} - t\delta\log p} \\ &\leq t\delta\frac{\log p}{6p^{1+\delta}}\end{aligned}$$

where the last inequality follows from $t\delta\log p \leq p^{1+\delta}$. We use a version of Chernoff's inequality for binomial distributions [41], which states that

$$\mathsf{P}(X > ap) \leq \left\{(q/a)^a \mathrm{e}^a\right\}^p \text{ if } X \sim \mathsf{Binom}(p, q) \text{ and } q \leq a < 1. \tag{3.7.1}$$

Since $\pi_\theta \le t\delta\frac{\log p}{6p^{1+\delta}} \le t/p$, by the above inequality, on the event $\mathcal{E}_k$, we have

$$\begin{aligned}
\Pi\left(S_k \ge t|(\theta_h)_{h\in[k]}\right) &= \Pi\left(S_k \ge \frac{t}{p}p\Big|(\theta_h)_{h\in[k]}\right) \\
&\le \left[\left\{\pi_\theta\frac{p}{t}\right\}^{t/p}\mathrm{e}^{t/p}\right]^p \\
&\le \left[\left\{\frac{t\delta\log p}{6p^{1+\delta}}\frac{p}{t}\right\}^{t/p}\mathrm{e}^{t/p}\right]^p \\
&\le \left\{\frac{\log p^{\delta/2}}{6p^\delta}\right\}^t \mathrm{e}^t \le \mathrm{e}^{-t(\delta/2)\log p}.
\end{aligned} \tag{3.7.2}$$

Now we will bound $\mathsf{P}(\mathcal{E}_k^c)$. Since $0 < \alpha < 1$, Gautschi's inequality implies

$$\mathsf{B}(\alpha,\kappa+1) = \frac{\Gamma(\alpha)}{\kappa+\alpha}\frac{\Gamma(\kappa+1)}{\Gamma(\kappa+\alpha)} > \frac{\Gamma(\alpha)}{\kappa+1}\kappa^{1-\alpha} \ge \frac{\kappa^{1-\alpha}}{\kappa+1}.$$

Let

$$t_0 := \frac{t\delta\log p}{7p^{1+\delta}}.$$

Then using the union bound, we have

$$\begin{aligned}
\Pi(\mathcal{E}_k^c) \le k\Pi\left(\nu_1 > t_0\right) &= \frac{k}{\mathsf{B}(\alpha,\kappa+1)}\int_{t_0}^1 \nu^{\alpha-1}(1-\nu)^\kappa \mathrm{d}\nu \\
&\le k\frac{\kappa+1}{\kappa^{1-\alpha}}t_0^{\alpha-1}\int_{t_0}^1(1-\nu)^\kappa\mathrm{d}\nu \\
&\le k\left(\frac{1}{\kappa t_0}\right)^{1-\alpha}\left(1-\frac{t\delta\log p}{7p^{1+\delta}}\right)^{p^{1+\delta}+1} \\
&\le k\exp\left((1-\alpha)\log(t\delta\log p/7)\right)\exp\left(-\frac{t\delta}{7}\log p\right).
\end{aligned}$$

Moreover, by the fact that $\log x \leq (1/\mathrm{e})x$ for any $x > 0$, we have that

$$\begin{aligned}
\mathsf{P}(\mathcal{E}^c) &\leq k \exp\left((1-\alpha)\frac{t\delta \log p}{7\mathrm{e}}\right) \exp\left(-\frac{1}{7}t\delta \log p\right) \\
&\leq k \exp\left(\frac{1}{7\mathrm{e}}t\delta \log p - \frac{1}{7}t\delta \log p\right) \\
&\leq k \exp\left(-\frac{(\mathrm{e}-1)}{7\mathrm{e}}t\delta \log p\right) \\
&\leq k \exp\left(-\frac{1}{12}t\delta \log p\right).
\end{aligned} \tag{3.7.3}$$

Combining (3.7.2) and (3.7.3) we obtain the desired bound. □

Proof of Lemma 3.2.3. Let $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_{0n,k}$ be the k-th columns of $\mathbf{B}$ and $\mathbf{B}_{0n}$, respectively. Since $\eta_n^2 = (6/\pi^2)\sum_{k=1}^{\infty}(\eta_n^2/k^2)$, it follows from Lemma 3.7.3 that

$$\begin{aligned}
&\mathsf{P}(\|\mathbf{B} - \mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n | \boldsymbol{\theta}) \\
&\geq \prod_{k=1}^{\infty} \mathsf{P}\left(\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0n,k}\|_2 \leq \frac{\sqrt{6}\eta_n}{\pi k} \middle| \theta_k\right) \\
&\geq \prod_{k=1}^{k_{0n}} \left[\theta_k^{s_n}(1-\theta_k)^{p_n - s_n} \mathrm{e}^{-\|\boldsymbol{\beta}_{0n,k}\|_1 - \sqrt{3}\eta_n/(\pi k)} \left(\frac{\sqrt{3}\eta_n}{\pi s_n k_{0n}}\right)^{s_n}\right] \prod_{k=k_0+1}^{\infty}(1-\theta_k)^{p_n} \\
&\geq \mathrm{e}^{-\|\mathbf{B}_{0n}\|_1 - \eta_n \sum_{k=1}^{k_{0n}} k^{-1}} \left(\frac{\eta_n}{2 s_n k_{0n}}\right)^{s_n k_{0n}} \prod_{k=1}^{k_0} \theta_k^s (1-\theta_k)^{p_n - s_n} \prod_{k=k_0+1}^{\infty}(1-\theta_k)^{p_n}.
\end{aligned}$$

Since $s_n k_{0n}^{-s_n k_{0n}} = \mathrm{e}^{-s_n k_{0n} \log(s_n k_{0n})} \geq \mathrm{e}^{-2 s_n k_{0n} \log p_n}$ and that $\sum_{k=1}^{k_{0n}} 1/k \leq \log(k_{0n}+1) \leq k_{0n}$, we further have

$$\begin{aligned}
&\mathsf{P}(\|\mathbf{B} - \mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n | \boldsymbol{\theta}) \\
&\geq \mathrm{e}^{-\|\mathbf{B}_{0n}\|_1 - \eta_n k_{0n} - 2 s_n k_{0n} \log(p_n \vee \eta_n^{-1})} \prod_{k=1}^{k_0} \theta_k^s (1-\theta_k)^{p_n - s_n} \prod_{k=k_0+1}^{\infty}(1-\theta_k)^{p_n}.
\end{aligned}$$

Since $\theta_1 > \theta_2 > \cdots$, we have

$$\prod_{k=1}^{k_{0n}} \theta_k^s (1-\theta_k)^{p_n - s_n} \geq \theta_{k_{0n}}^{s_n k_{0n}} (1-\theta_1)^{(p_n - s_n)k_{0n}}$$
$$= \left(\prod_{k=1}^{k_{0n}} \nu_k \right)^{s_n k_{0n}} (1-\nu_1)^{(p_n - s_n)k_{0n}}$$

Note that $(1-\theta_k) \geq (1 - \theta_k/\theta_{k_{0n}})$ and $\theta_k/\theta_{k_{0n}} = \prod_{h=k_{0n}+1}^{k} v_h$ is independent to $\{\theta_1, \ldots, \theta_{k_{0n}}\}$. Since $(\nu_h)_{h\in\mathbb{N}}$ are iid, we can see that $\{\theta_k/\theta_{k_{0n}} : k = k_{0n}+1, k_{0n}+2, \ldots\}$ as a lagged stick-breaking process which has the same distribution as $\{\theta_k : k \in \mathbb{N}\}$. Thus we have

$$\mathsf{E}\left[\prod_{k=1}^{k_{0n}} \theta_k^s (1-\theta_k)^{p_n - s_n} \prod_{k=k_{0n}+1}^{\infty} (1-\theta_k)^{p_n} \right]$$
$$\geq \left[\prod_{k=2}^{k_{0n}} \mathsf{E}\left(\nu_k^{s_n k_{0n}} \right) \right] \mathsf{E}\left\{ \nu_1^{s_n k_{0n}} (1-\nu_1)^{(p_n - s_n)k_{0n}} \right\}$$
$$\times \mathsf{E}\left\{ \prod_{k=k_{0n}+1}^{\infty} \left(1 - \theta_k/\theta_{k_{0n}}\right)^{p_n} \right\}$$
$$= \left\{ \mathsf{E}\left(\nu_k^{s_n k_{0n}} \right) \right\}^{k_{0n}-1} \mathsf{E}\left\{ \nu_1^{s_n k_{0n}} (1-\nu_1)^{(p_n - s_n)k_{0n}} \right\}$$
$$\times \mathsf{E}\left\{ \prod_{k=1}^{\infty} (1-\theta_k)^{p_n} \right\}.$$

Since $\mathsf{B}(\alpha_n, p_n^{1+\delta}+1) < \mathsf{B}(\alpha_n, 1) = \alpha_n^{-1}$ and $p_n^{1+\delta} > p_n - s_n$, it follows that

$$\begin{aligned}
&\mathsf{E}\left\{\nu_1^{s_n k_{0n}}(1-\nu_1)^{(p_n-s_n)k_{0n}}\right\} \\
&= \frac{1}{\mathsf{B}(\alpha_n, p_n^{1+\delta}+1)} \int_0^1 \nu^{s_n k_{0n}+\alpha_n-1}(1-\nu)^{p_n^{1+\delta}+(p_n-s_n)k_{0n}} \mathrm{d}\nu \\
&\geq \alpha_n \int_0^{p_n^{-(1+\delta)}} \nu^{s_n k_{0n}}(1-\nu)^{p_n^{1+\delta}(1+k_{0n})} \mathrm{d}\nu \\
&\geq \alpha_n \int_0^{p_n^{-(1+\delta)}} \nu^{s_n k_{0n}} \mathrm{d}\nu \left(1 - \frac{1}{p_n^{1+\delta}}\right)^{p_n^{1+\delta}(1+k_{0n})} \\
&\geq \frac{\alpha_n}{s_n k_{0n}+1} \left(\frac{1}{p_n^{1+\delta}}\right)^{s_n k_{0n}} \mathrm{e}^{-2(1+k_{0n})} \\
&\geq \alpha_n \mathrm{e}^{-C_1 s_n k_{0n} \log p_n},
\end{aligned}$$

for some $C_1 > 0$ depending only on δ, where the third inequality is due to the inequality $(1-x)^{1/x} \geq \mathrm{e}^{-2}$ for $0 < x < 1/2$. Similarly we have

$$\begin{aligned}
\mathsf{E}\left(\nu_k^{s_n k_{0n}}\right) &= \frac{1}{\mathsf{B}(\alpha_n, p_n^{1+\delta}+1)} \int_0^1 \nu^{s_n k_{0n}+\alpha_n-1} \mathrm{d}\nu \\
&\geq \alpha_n \mathrm{e}^{-C_2 s_n k_{0n} \log p_n}
\end{aligned}$$

for some $C_2 > 0$ depending only on δ. Lemma 3.7.1 implies that

$$\mathsf{E}\left\{\prod_{k=1}^{\infty}(1-\theta_k)\right\} \geq \exp\left\{-2\left(p_n \frac{\alpha_n+\kappa+1}{\kappa+1} + \frac{4}{3}\right)\left(\frac{\alpha_n}{\alpha+\kappa+1}\right)\right\} \gtrsim 1.$$

To combine all the results, we obtain

$$\begin{aligned}
\mathsf{P}(\|\mathbf{B}-\mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n | \boldsymbol{\theta}) &= \mathsf{E}\left[\mathsf{P}(\|\mathbf{B}-\mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n | \boldsymbol{\theta})\right] \\
&\geq \mathrm{e}^{-\|\mathbf{B}_{0n}\|_1 - \eta k_{0n}} \alpha_n^{k_{0n}} \mathrm{e}^{-C_3 s_n k_{0n} \log(p_n \vee \eta_n^{-1})}
\end{aligned}$$

for some $C_3 > 0$ depending only on δ, which is the desired result.

For the second assertion, note that (A2) implies

$$\|\mathbf{B}_{0n}\|_1 \leq \sqrt{s_n k_{0n}}\, \|\mathbf{B}_{0n}\|_{\mathrm{F}} \leq \sqrt{s_n} k_{0n} \|\mathbf{B}_{0n}\| \lesssim \sqrt{s_n c_n} k_{0n} \lesssim s_n k_{0n}$$

which completes the proof. □

Proof of Corollary 3.2.4. Let $\eta_n := \sqrt{s_n k_{0n}/n}$. Since $\mathbf{B}$ and σ^2 are independent,

$$\begin{aligned}&\Pi_n(\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\|_{\mathrm{F}} \le \eta_n) \\ &\ge \Pi_n\left(\left\|\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right\|_{\mathrm{F}} \le \frac{\eta_n}{2}\right)\Pi_n\left(\sqrt{p_n}|\sigma^2 - \sigma_{0n}^2| \le \frac{\eta_n}{2}\right).\end{aligned}$$

Since $\sigma_{0n}^2 \ge c_0 > 0$ by (A3) and $s_n k_{0n}/n < p_n$ by (A4), it follows that

$$\begin{aligned}\Pi_n\left(|\sigma^2 - \sigma_{0n}^2| \le \frac{\eta_n}{2\sqrt{p_n}}\right) &\ge \frac{\eta_n}{2\sqrt{p_n}} \min_{\sigma^2 \in [\sigma_{0n}^2/2, 3\sigma_{0n}^2/2]} \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}\mathrm{e}^{-b/\sigma^2} \\ &\ge C_1 p_n^{-1} c_n^{-a-1}\mathrm{e}^{-2b/c_0} \ge \mathrm{e}^{-C_2 \log p_n},\end{aligned}$$

where the positive constants C_1 and C_2 depend only on a, b and c_0.

We slightly abuse the notation to let $\mathbf{B}_{0n}$ be a $p \times \infty$ matrix where all columns after the k_{0n}-th column are $\mathbf{0}$. Since $\|\mathbf{B}_{0n}\| \lesssim \sqrt{c_n} \lesssim \sqrt{s_n}$ by (A2) and

$$\begin{aligned}\left\|\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right\|_{\mathrm{F}} &\le \|\mathbf{B}_{0n} - \mathbf{B}\|_{\mathrm{F}}^2 + 2\left\|\mathbf{B}_{0n}(\mathbf{B}_{0n} - \mathbf{B})^\top\right\|_{\mathrm{F}} \\ &\le \|\mathbf{B}_{0n} - \mathbf{B}\|_{\mathrm{F}}^2 + 2\|\mathbf{B}_{0n}\|\,\|\mathbf{B}_{0n} - \mathbf{B}\|_{\mathrm{F}},\end{aligned}$$

we have that $\|\mathbf{B}_{0n} - \mathbf{B}\|_{\mathrm{F}} \le C_3\sqrt{k_{0n}/n}$ implies $\left\|\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right\|_{\mathrm{F}} \le \eta_n/2$ for some $C_3 > 0$. Then Lemma 3.2.3, which shows

$$\Pi_n\left(\|\mathbf{B} - \mathbf{B}_{0n}\|_{\mathrm{F}} \le C_3\sqrt{\frac{k_{0n}}{n}}\right) \ge \alpha_n^{k_0 n}\mathrm{e}^{-C_4 s_n k_{0n}\log p_n}$$

for some universal constant $C_4 > 0$, and the fact that $n/k_{0n} \lesssim p_n$ by (A4) complete the proof. □

3.7.2 Proofs of theorems in Section 3.3

To simplify notation, we write $\mathsf{P}_0 := \mathsf{P}_{\boldsymbol{\Sigma}_{0n}}$ and $\mathsf{E}_0 := \mathsf{E}_{\boldsymbol{\Sigma}_{0n}}$.

Proof of Theorem 3.3.1. Fix $\mathbf{\Sigma}_{0n} \in \mathcal{C}_{0n}$. For a Borel measurable subset B of the parameter space, the posterior probability of B is written as

$$\Pi_n\left(B|\mathbf{Y}_{1:n}\right) = \frac{\int_B \prod_{i=1}^n \frac{f_{\mathbf{\Sigma}}(\mathbf{Y}_i)}{f_{\mathbf{\Sigma}_{0n}}(\mathbf{Y}_i)} \mathrm{d}\Pi_n(\mathbf{\Sigma})}{\int \prod_{i=1}^n \frac{f_{\mathbf{\Sigma}}(\mathbf{Y}_i)}{f_{\mathbf{\Sigma}_{0n}}(\mathbf{Y}_i)} \mathrm{d}\Pi_n(\mathbf{\Sigma})} \equiv \frac{N_n(B)}{D_n},$$

where $f_{\mathbf{\Sigma}}$ denotes the density of $\mathsf{N}(\mathbf{0}, \mathbf{\Sigma})$ and $N_n(B)$ and D_n denote the numerator and denominator of the fraction in the preceding display. By Corollary 3.2.4 and Lemma 3.7.4, we have that there is a Borel measurable set $\mathfrak{A}_n \in \sigma(\mathbf{Y}_{1:n})$ with

$$\mathsf{P}_0(\mathfrak{A}_n) \leq C_1 / \log n \tag{3.7.4}$$

for some universal constant $C_1 > 0$, on which

$$D_n \geq \alpha_n^{k_0 n} \mathrm{e}^{-C_2 s_n k_{0n} \log p_n} \equiv \alpha_n^{k_{0n}} \mathrm{e}^{-r_n}, \tag{3.7.5}$$

where

$$r_n := C_2 s_n k_{0n} \log p_n.$$

We define the following two sets

$$\mathcal{F}_{n,1} := \left\{\mathbf{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2 \mathbf{I} : \mathbf{B}_{>u_n} = \mathbf{0}\right\}$$
$$\mathcal{F}_{n,2} := \left\{\mathbf{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2 \mathbf{I} : |\mathsf{supp}_{\lfloor u_n \rfloor}(\mathbf{B})| \leq t_n\right\},$$

where t_n and u_n are sequences that will be specified later. Let

$$\mathcal{U}_n := \left\{\mathbf{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2 \mathbf{I} : \|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| \geq M\epsilon_n\right\}.$$

We decompose the posterior probability as

$$\Pi_n(\mathcal{U}_n|\mathbf{Y}_{1:n}) \leq \Pi_n(\mathcal{U}_n^\star|\mathbf{Y}) + \Pi_n(\mathcal{F}_{n,2}^c|\mathbf{Y}_{1:n}) + \Pi_n(\mathcal{F}_{n,1}^c|\mathbf{Y}_{1:n}),$$

where $\mathcal{U}_n^\star := \mathcal{U}_n \cap \mathcal{F}_{n,2} \cap \mathcal{F}_{n,1}$. By Lemma 3.2.1, we have that

$$\begin{aligned}\Pi_n(\mathcal{F}_{n,1}^c) &= 1 - \Pi_n(\mathbf{B}_{>u_n} = \mathbf{0}) \\ &\leq 6p_n \left(\frac{\alpha_n}{\alpha_n + p_n^{1+\delta} + 1}\right)^{u_n+1} \\ &\leq 6\left(\frac{\alpha_n}{p_n}\right)^{u_n} \leq 6\left(\frac{1}{p_n}\right)^{u_n}.\end{aligned}$$

If we take

$$u_n := \frac{1}{\log p_n}\left\{2r_n + \log\left(\frac{1}{\alpha_n}\right)\right\}$$

then

$$\begin{aligned}\mathsf{E}_0[\Pi_n(\mathcal{F}_{n,1}^c|\mathbf{Y}_{1:n})] &\leq \mathsf{E}_0\left[\frac{\Pi_n(\mathcal{F}_{n,1}^c)}{D_n}\mathbb{1}_{\mathfrak{A}_n}\right] + \mathsf{P}_0(\mathfrak{A}_n^c) \\ &\leq \alpha_n^{-k_{0n}}\mathrm{e}^{r_n}\Pi_n(\mathcal{F}_{n,1}^c) + \frac{C_1}{\log n} \\ &\leq \mathrm{e}^{-r_n} + \frac{C_1}{\log n}.\end{aligned}$$

We invoke Lemma 3.2.2 with the inequality $x + 1 \leq \mathrm{e}^x$ to obtain

$$\begin{aligned}\mathsf{E}_0\left[\Pi_n(\mathcal{F}_{n,2}^c|\mathbf{Y}_{1:n})\right] &\leq \alpha_n^{-k_{0n}}\mathrm{e}^{r_n}\Pi_n(\mathcal{F}_{n,2}^c) + \frac{C_1}{\log n} \\ &\leq (u_n+1)\alpha_n^{-k_{0n}}\mathrm{e}^{r_n}\mathrm{e}^{-C_4 t_n \log p_n} + \frac{C_1}{\log n} \\ &\leq \alpha_n^{-2k_{0n}}\mathrm{e}^{3r_n}\mathrm{e}^{-C_4 t_n \log p_n} + \frac{C_1}{\log n}\end{aligned}$$

for some universal constant $C_4 > 0$. Hence if we let

$$\begin{aligned}\tau_n &:= k_{0n}\max\left\{s_n \log p_n, \log\left(\frac{1}{\alpha_n}\right)\right\} \\ t_n &:= C_5\tau_n/\log p_n\end{aligned}$$

for sufficiently large $C_5 > 0$, then we have that $\mathsf{E}_0[\Pi(\mathcal{F}_{n,2}^c|\mathbf{Y}_{1:n})] \lesssim \mathrm{e}^{-C_6\tau_n} + 1/\log n \lesssim 1/\log n$ for some universal constant $C_6 > 0$.

For $\mathcal{U}_n^\star$, note that for any test function ϕ_n, and $\mathfrak{A}_n \in \sigma(\mathbf{Y}_{1:n})$,

$$\mathsf{E}_0\left[\Pi_n\left(\mathcal{U}_n^\star|\mathbf{Y}_{1:n}\right)\right] \leq \mathsf{E}_0\phi_n + \mathsf{E}_0\left[(1-\phi_n)\Pi_n\left(\mathcal{U}_n^\star|\mathbf{Y}_{1:n}\right)\mathbb{1}_{\mathfrak{A}_n}\right] + \mathsf{P}_0(\mathfrak{A}_n^c).$$

Suppose $\mathfrak{A}_n$ satisfies (3.7.4) and (3.7.5), then we have

$$\begin{aligned}
&\mathsf{E}_0\left[(1-\phi_n)\Pi(\mathcal{U}_n^\star|\mathbf{Y}_{1:n})\mathbb{1}_{\mathfrak{A}_n}\right] \\
&\leq \mathsf{E}_0\left[(1-\phi_n)\frac{N_n(\mathcal{U}_n^\star)}{D_n}\mathbb{1}_{\mathfrak{A}_n}\right] \\
&\leq \frac{1}{\alpha_n^{k_{0n}}\mathrm{e}^{-r_n}}\mathsf{E}_0\left[(1-\phi_n)\int_{\mathcal{U}_n^\star}\prod_{i=1}^n\frac{f_{\boldsymbol{\Sigma}}(\mathbf{Y}_i)}{f_{\boldsymbol{\Sigma}_{0n}}(\mathbf{Y}_i)}\mathrm{d}\Pi_n(\boldsymbol{\Sigma})\right] \\
&= \frac{1}{\alpha_n^{k_{0n}}\mathrm{e}^{-r_n}}\int\left[(1-\phi_n)\int_{\mathcal{U}_n^\star}\prod_{i=1}^n\frac{f_{\boldsymbol{\Sigma}}(\mathbf{Y}_i)}{f_{\boldsymbol{\Sigma}_{0n}}(\mathbf{Y}_i)}\mathrm{d}\Pi_n(\boldsymbol{\Sigma})\prod_{i=1}^n f_{\boldsymbol{\Sigma}_{0n}}(\mathbf{Y}_i)\mathrm{d}\mathbf{Y}_i\right] \\
&\leq \frac{1}{\alpha_n^{k_{0n}}\mathrm{e}^{-r_n}}\int_{\mathcal{U}_n^\star}\mathsf{E}_{\boldsymbol{\Sigma}}(1-\phi_n)\mathrm{d}\Pi_n(\boldsymbol{\Sigma}) \\
&\leq \frac{1}{\alpha_n^{k_{0n}}\mathrm{e}^{-r_n}}\sup_{\boldsymbol{\Sigma}\in\mathcal{U}_n^\star}\mathsf{E}_{\boldsymbol{\Sigma}}(1-\phi_n)
\end{aligned}$$

and hence

$$\mathsf{E}_0\left[\Pi_n\left(\mathcal{U}_n^\star|\mathbf{Y}_{1:n}\right)\right] \leq \mathsf{E}_0\phi_n + \alpha_n^{-k_{0n}}\mathrm{e}^{r_n}\sup_{\boldsymbol{\Sigma}\in\mathcal{U}_n^\star}\mathsf{E}_{\boldsymbol{\Sigma}}(1-\phi_n) + \frac{C_1}{\log n}.$$

By Lemma 3.7.5, there is a test function ϕ_n such that

$$\mathsf{E}_0\phi_n \leq 3\exp\left(C_7 t_n\log p_n - C_8 M^{1/2}\frac{n\epsilon_n^2}{c_n^2}\right)$$

$$\alpha_n^{-k_{0n}}\mathrm{e}^{r_n}\sup_{\boldsymbol{\Sigma}\in\mathcal{U}_n^\star}\mathsf{E}_{\boldsymbol{\Sigma}}(1-\phi_n) \leq \exp\left(k_{0n}\log\left(\frac{1}{\alpha_n}\right) + r_n + C_9(t_n+s_n) - C_{10}Mn\epsilon_n^2\right)$$

for some universal positive constants $C_7, \ldots, C_{10}$. Since $k_{0n} \log(1/\alpha_n) \lesssim \tau_n$, $r_n \lesssim \tau_n$, and $t_n + s_n \lesssim \tau_n$, we have

$$\begin{aligned}
&\mathsf{E}_0 \phi_n + \alpha_n^{-k_{0n}} \mathrm{e}^{r_n} \sup_{\boldsymbol{\Sigma} \in \mathcal{U}_n^\star} \mathsf{E}_{\boldsymbol{\Sigma}} (1 - \phi_n) \\
&\lesssim \exp\left(C_{12} \tau_n - C_8 M^{1/2} \frac{n \epsilon_n^2}{c_n^2} \right) + \exp\left(C_{11} \tau_n - C_{10} M n \epsilon_n^2 \right) \\
&\lesssim \exp\left((C_{13} - C_{14} M^{1/2}) k_{0n} \max\left\{ s_n \log p_n, \log\left(\frac{1}{\alpha_n} \right) \right\} \right)
\end{aligned}$$

for some universal positive constants $C_{10}, \ldots, C_{14}$. Hence for sufficiently large M such that $M > C_{13}^2 / C_{14}^2$, we obtain the desired result. □

Proof of Theorem 3.3.2. Fix $\boldsymbol{\Sigma}_{0n} \in \mathcal{C}_{0n}^\star$. Let $\alpha_n = p_n^{-A s_n^2}$. By Corollary 3.2.4 and Lemma 3.7.4, we have that there is a Borel measurable set $\mathfrak{A}_n$ with $\mathsf{P}_0(\mathfrak{A}_n) \leq C_1 / \log n$ for some universal constant $C_1 > 0$, on which

$$D_n := \int \prod_{i=1}^n \frac{f_{\boldsymbol{\Sigma}}(\mathbf{Y}_i)}{f_{\boldsymbol{\Sigma}_{0n}}(\mathbf{Y}_i)} \mathrm{d}\Pi_n(\boldsymbol{\Sigma}) \geq \alpha_n^{k_{0n}} \mathrm{e}^{-r_n},$$

where

$$r_n := C_2 s_n k_{0n} \log p_n$$

for some universal constant $C_2 > 0$. Since

$$\Pi_n\left(\mathsf{K}^+(\mathbf{B}) > k_{0n} \right) \leq 6 p_n \left(\frac{\alpha_n}{\alpha_n + p_n^{1+\delta} + 1} \right)^{k_{0n}+1} \leq 4 \alpha_n^{k_{0n}+1},$$

we have that

$$\begin{aligned}\mathsf{E}_0\left[\Pi_n\left(\mathsf{K}^+(\mathbf{B}) > k_{0n} | \mathbf{Y}_{1:n}\right)\right] &\leq \mathsf{E}_0\left[\Pi_n\left(\mathsf{K}^+(\mathbf{B}) > k_{0n}\right)\mathbb{1}_{\mathfrak{A}_n}\right] + \mathsf{P}_0(\mathfrak{A}_n^c) \\ &\leq \mathrm{e}^{r_n}\alpha_n^{-k_{0n}}\Pi_n\left(\mathsf{K}^+(\mathbf{B}) > k_{0n}\right) + \frac{C_4}{\log n} \\ &\leq 4\mathrm{e}^{r_n}\alpha_n + \frac{C_4}{\log n} \\ &\leq 4\mathrm{e}^{C_2 s_n k_{0n}\log p_n - A s_n^2 \log p_n} + \frac{C_4}{\log n}.\end{aligned} \tag{3.7.6}$$

Since $s_n \geq k_{0n}$ by the assumption (A5), if A is larger than C_2, the posterior probability of the event $\left\{\mathsf{K}^+(\mathbf{B}) > k_{0n}\right\}$ converges to zero as n goes to infinity.

For the set $\left\{\mathsf{K}^+(\mathbf{B}) < k_{0n}\right\}$, by Theorem 3.3.1, we have

$$\mathsf{E}_0\left[\Pi_n\left(\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\| > M\epsilon_n \middle| \mathbf{Y}_{1:n}\right)\right] = o(1) \tag{3.7.7}$$

for sufficiently large $M > 0$, where

$$\epsilon_n := c_n\sqrt{\frac{s_n^2 k_{0n}\log p_n}{n}}. \tag{3.7.8}$$

In addition, by (A4), the contraction rate ϵ_n goes to zero. Suppose that $\mathsf{K}^+(\mathbf{B}) < k_{0n}$. Since $\mathbf{B}_{0n}$ is of full rank, there is $\mathbf{v}_1 \in \mathsf{span}(\mathbf{B})^\perp \cap \mathsf{span}(\mathbf{B}_{0n})$ with $\|\mathbf{v}_1\|_2 = 1$. Then by (A5),

$$\left\|\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right\| \geq \left\|\left(\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right)\mathbf{v}_1\right\|_2 = \left\|\mathbf{B}_{0n}\mathbf{B}_{0n}^\top\mathbf{v}_1\right\|_2 > d_0.$$

Let $\mathcal{B} := \mathsf{span}(\mathbf{B}) \cup \mathsf{span}(\mathbf{B}_{0n})$. Since $\mathsf{rank}(\mathcal{B}) < 2k_{0n} < p_n$, there is $\mathbf{v}_2 \in \mathcal{B}^\perp$ with $\|\mathbf{v}_2\|_2 = 1$. Then $\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\| \geq \left\|\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\right)\mathbf{v}_2\right\|_2 = \left|\sigma^2 - \sigma_{0n}^2\right|$. Hence by the triangular inequality,

$$\left\|\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right\| \leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\| + \left\|\left(\sigma^2 - \sigma_{0n}^2\right)\mathbf{I}\right\| \leq 2\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\|. \tag{3.7.9}$$

Since ϵ_n goes to zero,

$$\begin{aligned}\mathsf{E}_0\left[\Pi_n\left(\mathsf{K}^+(\mathbf{B}) < k_{0n}\middle|\mathbf{Y}_{1:n}\right)\right] &\leq \mathsf{E}_0\left[\Pi_n\left(\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0n}\| > d_0/2\middle|\mathbf{Y}_{1:n}\right)\right] \\ &\leq \mathsf{E}_0\left[\Pi_n\left(\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0n}\| > M\epsilon_n\middle|\mathbf{Y}_{1:n}\right)\right]\end{aligned}$$

as n goes to infinity, which completes the proof. □

3.7.3 Proof of Theorem 3.5.1

Proof of Theorem 3.5.1. To simplify notation, we write $\mathsf{E}_0 := \mathsf{E}_{\boldsymbol{\Sigma}_{0n}}$. Fix $\boldsymbol{\Sigma}_{0n} \in \mathcal{C}_{0n}^{\star\star}$. By Lemma 3.7.4, there is a Borel measurable set $\mathfrak{A}_n \in \sigma(\mathbf{Y}_{1:n})$ with $\mathsf{P}_0(\mathfrak{A}_n) \lesssim 1/\log n$ on which

$$\begin{aligned}D_n &:= \int \prod_{i=1}^{n} \frac{f_{\boldsymbol{\Sigma}}(\mathbf{Y}_i)}{f_{\boldsymbol{\Sigma}_{0n}}(\mathbf{Y}_i)} \mathrm{d}\Pi_n(\boldsymbol{\Sigma}) \\ &\geq \mathrm{e}^{-C_3 s_n k_{0n} \log s_n} \check{\Pi}_n\left(\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0n}\|_{\mathrm{F}} < \sqrt{s_n k_{0n}/n}\right).\end{aligned}$$

Note that

$$\begin{aligned}&\check{\Pi}_n\left(\|\mathbf{B}-\mathbf{B}_{0n}\|_{\mathrm{F}} \geq \sqrt{k_{0n}/n}\right) \\ &\geq \check{\Pi}_{n,1}(s_n, k_{0n}) \frac{1}{\binom{p_n}{s_n}} \check{\Pi}_n\left(\left\|\mathbf{B}^{S_0}_{\leq k_{0n}} - \mathbf{B}^{S_0}_{0n,\leq k_{0n}}\right\|_{\mathrm{F}} \geq \sqrt{k_{0n}/n}\right).\end{aligned}$$

By Lemma 3.7.2 and the assumption (A2) which implies $\|\mathbf{B}^{S_0}_{0n,\leq k_{0n}}\|_1 \lesssim s_n k_{0n}$, we have that

$$\check{\Pi}_n\left(\left\|\mathbf{B}^{S_0}_{\leq k_{0n}} - \mathbf{B}^{S_0}_{0n,\leq k_{0n}}\right\|_{\mathrm{F}} \geq \sqrt{k_{0n}/n}\right) \geq \mathrm{e}^{-C_1 s_n k_{0n} \log n}$$

for some universal constant $C_1 > 0$. Thus

$$\begin{aligned}
&\check{\Pi}_n\left(\|\mathbf{B}-\mathbf{B}_{0n}\|_{\mathrm{F}} \geq \sqrt{k_{0n}/n}\right) \\
&\geq \exp(-As_n^2 k_{0n}\log p_n - s_n \log p_n - C_1 s_n k_{0n}\log n) \\
&\geq \exp(-As_n^2 k_{0n}\log p_n - C_2 s_n k_{0n}\log p_n)
\end{aligned}$$

for some universal constant $C_2 > 0$. By using similar arguments used in the proof of Corollary 3.2.4,

$$D_n \geq \exp\left(-C_3 s_n k_{0n}\log p_n - As_n^2 k_{0n}\log p_n\right)$$

for some universal constant $C_3 > 0$ on $\mathfrak{A}_n$

For the posterior contraction of the covariance matrix, we let

$$u_n := s_n^2 k_{0n} + C_3 s_n k_{0n}/A$$

Define

$$\mathcal{F}_{n,1} := \left\{\mathsf{K}^+(\mathbf{B}) \leq u_n\right\}, \quad \mathcal{F}_{n,2} := \left\{|\mathsf{supp}_{\lfloor u_n \rfloor}(\mathbf{B})| \leq \sqrt{u_n}\right\}.$$

Then

$$\check{\Pi}_n(\mathcal{F}_{n,1}^c) \lesssim \sum_{s=1}^{p_n}\sum_{k=\lceil u_n\rceil}^{\infty} p_n^{-As^2k} \lesssim p_n^{-Au_n}$$

and

$$\check{\Pi}_n(\mathcal{F}_{n,2}^c) \lesssim \sum_{s=\lceil u_n\rceil}^{p_n}\sum_{k=1}^{\infty} p_n^{-As^2k} \leq p_n^{-Au_n}.$$

Hence both $\mathsf{E}_0\left[\check{\Pi}_n\left(\mathcal{F}_{n,1}^c|\mathbf{Y}_{1:n}\right)\right]$ and $\mathsf{E}_0\left[\check{\Pi}_n\left(\mathcal{F}_{n,2}^c|\mathbf{Y}_{1:n}\right)\right]$ converge to zero as $n \to \infty$.

Therefore it suffices to show that $\mathsf{E}_0\left[\Pi_n(\mathcal{U}_n^\star|\mathbf{Y}_{1:n})\right]$ converges to zero as $n \to \infty$, where

$$\mathcal{U}_n^\star := \left\{\mathbf{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2\mathbf{I} : \|\mathbf{\Sigma}-\mathbf{\Sigma}_{0n}\| \geq M\epsilon_n, |\mathsf{supp}_{\lfloor u_n\rfloor}(\mathbf{B})| \leq \sqrt{u_n}, \mathsf{K}^+(\mathbf{B}) \leq u_n\right\}.$$

with

$$\epsilon_n := c_n\sqrt{\frac{s_n^2 k_{0n}\log p_n}{n}}.$$

Then Lemma 3.7.5 implies that there is a test function ϕ_n such that

$$\begin{aligned}
\mathsf{E}_0\left[\check{\Pi}_n\left(\mathcal{U}_n^\star|\mathbf{Y}_{1:n}\right)\right] &\le \mathsf{E}_0\phi_n + \mathsf{E}_0\left[(1-\phi_n)\check{\Pi}_n\left(\mathcal{U}_n^\star|\mathbf{Y}_{1:n}\right)\mathbb{1}_{\mathfrak{A}_n}\right] + \mathsf{P}_0(\mathfrak{A}_n^c) \\
&\lesssim \mathsf{E}_0\phi_n + \mathrm{e}^{(A+C_3)s_n^2 k_{0n}\log p_n}\sup_{\boldsymbol{\Sigma}\in\mathcal{U}_n^\star}\mathsf{E}_{\boldsymbol{\Sigma}}(1-\phi_n) + \frac{1}{\log n} \\
&\lesssim \mathrm{e}^{C_4\sqrt{u_n}\log p_n - C_5 M^{1/2} n\epsilon_n^2/c_n^2} \\
&\quad + \mathrm{e}^{(A+C_3)s_n^2 k_{0n}\log p_n + C_6(\sqrt{u_n}+s_n) - C_7 M n\epsilon_n^2} + \frac{1}{\log n} \\
&\lesssim \exp\left((C_8 - C_9 M^{1/2})s_n^2 k_{0n}\log p_n\right) + \frac{1}{\log n}
\end{aligned}$$

for some universal positive constants $C_4, \ldots, C_9$. Hence for sufficiently large $M > 0$, it follows that

$$\mathsf{E}_0\left[\check{\Pi}_n\left(\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\| > M\epsilon_n|\mathbf{Y}_{1:n}\right)\right] = o(1). \tag{3.7.10}$$

For the posterior consistency of the factor dimensionality, by similar arguments used in the proof of Theorem 3.3.2, we can show that

$$\mathsf{E}_0\left[\check{\Pi}_n\left(\mathsf{K}^+(\mathbf{B}) < k_{0n}|\mathbf{Y}_{1:n}\right)\right] \le \mathsf{E}_0\left[\check{\Pi}_n\left(\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\| > M\epsilon_n|\mathbf{Y}_{1:n}\right)\right] = o(1).$$

For the event $\left\{\mathsf{K}^+(\mathbf{B}) > k_{0n}\right\}$, we decompose the posterior probability as

$$\check{\Pi}_n\left(\mathsf{K}^+(\mathbf{B}) > k_{0n}\middle|\mathbf{Y}_{1:n}\right) \le \check{\Pi}_n\left(\mathsf{K}^+(\mathbf{B}) > k_{0n}, |\mathsf{supp}_{k_{0n}}(\mathbf{B})| \ge s_n\middle|\mathbf{Y}_{1:n}\right) \tag{3.7.11}$$

$$+ \check{\Pi}_n\left(|\mathsf{supp}_{k_{0n}}(\mathbf{B})| < s_n\middle|\mathbf{Y}_{1:n}\right). \tag{3.7.12}$$

Note that

$$\check{\Pi}_n\left(K > k_{0n}, |\mathsf{supp}_{k_{0n}}(\mathbf{B})| \ge s_n\right) \lesssim \sum_{k=k_{0n}+1}^{\infty}\sum_{s=s_n}^{\infty} p_n^{-As^2k} \lesssim p_n^{-As_n^2(k_{0n}+1)}$$

Hence we have that (3.7.11) converges to zero as $n \to \infty$ when $A > C_3$. For (3.7.12), suppose that $|\mathsf{supp}_{k_{0n}}(\mathbf{B})| < s_n$. Then $S^- := \mathsf{supp}_{k_{0n}}(\mathbf{B}_{0n}) \setminus$

$\mathsf{supp}_{k_{0n}}(\mathbf{B})$ is not empty. Let $\mathbf{v} \equiv (v_j)_{j\in[p]}$ be the vector such that $v_{j^*} = 1$ for $j^* \in S^-$ and 0 otherwise. Then

$$\left\|\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right\| \geq \left\|\mathbf{B}_{0n}\mathbf{B}_{0n}^\top\mathbf{v}\right\|_2 \geq \sqrt{\sum_{k=1}^{k_{0n}} \beta_{0n,j^*k}^2} > b_0$$

where b_0 is the lower bound of the nonzero entries of $\mathbf{B}_{0n}$. Since

$$\left\|\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right\| \leq 2\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\|$$

by (3.7.9) in the proof of Theorem 3.3.2, we have

$$\{|\mathsf{supp}_{k_{0n}}(\mathbf{B})| < s_n\} \subset \{\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| \geq b_0/2\}$$

and hence

$$\begin{aligned}\mathsf{E}_0\left[\check{\Pi}_n\left(|S| < s_n \middle| \mathbf{Y}_{1:n}\right)\right] &\leq \mathsf{E}_0\left[\check{\Pi}_n\left(\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| > b_0/2 | \mathbf{Y}_{1:n}\right)\right] \\ &\leq \mathsf{E}_0\left[\check{\Pi}_n\left(\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| > M\epsilon_n | \mathbf{Y}_{1:n}\right)\right]\end{aligned}$$

for all sufficiently large n. Since the last term of the above inequality converges to zero as $n \to \infty$ by (3.7.10), we complete the proof. □

3.7.4 Auxiliary lemmas

Lemma 3.7.1. *Let $\theta_h := \prod_{l=1}^h \nu_l$ for $h \in \mathbb{N}$, where $\nu_l \overset{\text{iid}}{\sim} \mathsf{Beta}(\alpha, \kappa+1)$ for $l \in \mathbb{N}$. Then, for any $k \in \mathbb{N}$ and $p > 0$*

$$\mathsf{E}\left[\prod_{h=k}^{\infty}(1-\theta_h)^p\right] \geq \exp\left\{-2\left(p\frac{\alpha+\kappa+1}{\kappa+1} + \frac{4}{3}\right)\left(\frac{\alpha}{\alpha+\kappa+1}\right)^k\right\}.$$

Proof. Define $\mathcal{E}_k := \left\{(\theta_h)_{h\in\mathbb{N}} : \theta_h \leq 3/4, \forall h >\geq k\right\}$. Then we have

$$\begin{aligned}
\mathsf{E}\left[\prod_{h=k}^{\infty}(1-\theta_h)^p\right] &\geq \mathsf{P}(\mathcal{E}_k)\mathsf{E}\left[\prod_{h=k}^{\infty}(1-\theta_h)^p\middle|\mathcal{E}_k\right] \\
&\geq \mathsf{P}(\mathcal{E}_k)\mathsf{E}\left[\exp\left(-2p\sum_{h=k}^{\infty}\theta_h\right)\middle|\mathcal{E}_k\right] \\
&\geq \mathsf{P}(\mathcal{E}_k)\exp\left(-2p\sum_{h=k}^{\infty}\mathsf{E}\left[\theta_h|\mathcal{E}_k\right]\right).
\end{aligned}$$

where the last inequality is due to Jensen's inequality. Note that

$$\begin{aligned}
\sum_{h=k}^{\infty}\mathsf{E}\left[\theta_h|\mathcal{E}_k\right] &\leq \sum_{h=k}^{\infty}\mathsf{E}\left[\theta_h\right] \\
&= \sum_{h=k}^{\infty}\left(\frac{\alpha}{\alpha+\kappa+1}\right)^h \\
&= \frac{\alpha+\kappa+1}{\kappa+1}\left(\frac{\alpha}{\alpha+\kappa+1}\right)^k
\end{aligned}$$

Since $1 \geq \theta_1 \geq \theta_2 \geq \cdots$, we have $\mathsf{P}(\mathcal{E}_k) = \mathsf{P}(\theta_k \leq 3/4)$. By Markov's inequality

$$\begin{aligned}
\mathsf{P}\left(\theta_1 \leq \frac{3}{4}\right) &= 1 - \mathsf{P}\left(\theta_k > \frac{3}{4}\right) \\
&\geq 1 - \frac{4}{3}\left(\frac{\alpha}{\alpha+\kappa+1}\right)^K \\
&\geq \exp\left(-\frac{8}{3}\left(\frac{\alpha}{\alpha+\kappa+1}\right)^k\right),
\end{aligned}$$

where we use the inequality $1 - x > \mathrm{e}^{-2x}$ for any $x \in (0, 1.5)$. This completes the proof. □

Lemma 3.7.2. *Assume that* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s)^\top$ *is distributed as* $\beta_j \overset{\text{iid}}{\sim} \mathsf{Laplace}(1)$ *for* $j \in [s]$*. Then for any* $\boldsymbol{\beta}_0 \in \mathbb{R}^s$ *and any* $\epsilon > 0$*,*

$$\mathsf{P}(\left\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\right\|_2 \leq \epsilon) \geq \mathrm{e}^{-\left\|\boldsymbol{\beta}_0\right\|_1 - \epsilon - s\log(s/\epsilon)}.$$

Proof. Using a change of variables $\boldsymbol{\beta}^{S_0} - \boldsymbol{\beta}_0^{S_0} \to \boldsymbol{\beta}^*$, we get

$$\begin{aligned}
\mathsf{P}\left(\left\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\right\|_1 \leq \epsilon\right) &= \int_{\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\right\|_1 \leq \epsilon} \frac{1}{2}\mathrm{e}^{-\|\boldsymbol{\beta}\|_1}\mathrm{d}\boldsymbol{\beta} \\
&\geq \mathrm{e}^{-\left\|\boldsymbol{\beta}_0\right\|_1} \int_{\left\|\boldsymbol{\beta}^*\right\|_1 \leq \epsilon} \frac{1}{2}\mathrm{e}^{-\|\boldsymbol{\beta}^*\|_1}\mathrm{d}\boldsymbol{\beta}^* \\
&= \mathrm{e}^{-\left\|\boldsymbol{\beta}_0\right\|_1}\mathsf{P}\left(\sum_{i=1}^{s} E_i \leq \epsilon\right)
\end{aligned}$$

where $E_1, \ldots, E_s$ are iid exponential random variables with scale 1. Recall that the sum of n iid exponential random variables with scale θ follows gamma distribution with shape n and scale θ. Thus we have

$$\begin{aligned}
\mathsf{P}\left(\sum_{i=1}^{s} E_i \leq \epsilon\right) &= \frac{1}{\Gamma(s)}\int_0^\epsilon x^{s-1}\mathrm{e}^{-x}\mathrm{d}x \\
&\geq \frac{1}{(s-1)!}\mathrm{e}^{-\epsilon}\int_0^\epsilon x^{s-1}\mathrm{d}x = \frac{\epsilon^s}{s!}\mathrm{e}^{-\epsilon}.
\end{aligned}$$

The fact that $s! \leq s^s$ completes the proof. □

Lemma 3.7.3. *Assume that* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ *is distributed as*

$$\begin{aligned}
\beta_j | z_j &\overset{\text{ind}}{\sim} (1 - \xi_j)\delta_0 + \xi_j \mathsf{Laplace}(1) \\
\xi_j &\overset{\text{iid}}{\sim} \mathsf{Bernoulli}(\theta)
\end{aligned}$$

for $\theta \in (0, 1)$*. Assume that* $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ *has a nonzero support* $S_0 := \mathsf{supp}(\boldsymbol{\beta}_0)$*. Let* $s = |S_0|$*. Then for any* $\epsilon > 0$*,*

$$\mathsf{P}(\left\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\right\|_2 \leq \epsilon) \geq \theta^s(1-\theta)^{p-s}\left[\mathrm{e}^{-\left\|\boldsymbol{\beta}_0\right\|_1 - \epsilon/\sqrt{2}}\left(\frac{\epsilon}{\sqrt{2}s}\right)^s\right].$$

Proof. We start with the inequality

$$\mathsf{P}\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2 \leq \epsilon\right) \geq \left[\prod_{j\in S_0^c} \mathsf{P}\left(|\beta_j| \leq \frac{\epsilon^2}{p}\right)\right] \mathsf{P}\left(\left\|\boldsymbol{\beta}^{S_0}-\boldsymbol{\beta}_0^{S_0}\right\|_2^2 \leq \frac{\epsilon^2}{2}\right).$$

Note that $\mathsf{P}\left(|\beta_j| \leq \frac{\epsilon^2}{p}\right) \geq \mathsf{P}\left(\beta_j = 0\right) = 1-\theta$ and that

$$\begin{aligned}\mathsf{P}\left(\left\|\boldsymbol{\beta}^{S_0}-\boldsymbol{\beta}_0^{S_0}\right\|_2^2 \leq \frac{\epsilon^2}{2}\right) &\geq \theta^s \mathsf{P}_{\mathsf{Lap}}\left(\left\|\boldsymbol{\beta}^{S_0}-\boldsymbol{\beta}_0^{S_0}\right\|_2^2 \leq \frac{\epsilon^2}{2}\right)\\ &\geq \theta^s \mathsf{P}_{\mathsf{Lap}}\left(\left\|\boldsymbol{\beta}^{S_0}-\boldsymbol{\beta}_0^{S_0}\right\|_1 \leq \frac{\epsilon}{\sqrt{2}}\right)\end{aligned}$$

where $\mathsf{P}_{\mathsf{Lap}}$ denotes the probability measure under the product of $\mathsf{Laplace}(1)$ densities. Lemma 3.7.2 completes the proof. □

Lemma 3.7.4 (Lemma 9.1 of [75]). *Let $\boldsymbol{\Sigma}_{0n}$ be a $p_n \times p_n$ symmetric positive definite matrix. Let η_n be a sequence satisfying $\eta_n^2/\lambda_{p_n}(\boldsymbol{\Sigma}_{0n}) \to 0$ and $n\eta_n^2/s^2_{\min}(\boldsymbol{\Sigma}_{0n}) \to \infty$, and let $\rho_n = 2\lambda_1(\boldsymbol{\Sigma}_{0n})/\lambda_{p_n}(\boldsymbol{\Sigma}_{0n})$. Then for any sequence of prior distributions $(\Pi_n(\cdot))_{n\in\mathbb{N}}$, there exists a Borel measurable set $\mathfrak{A}_n \in \sigma(\mathbf{Y}_{1:n})$ with $\mathsf{P}_{\boldsymbol{\Sigma}_{0n}}(\mathfrak{A}_n) \leq C_1/\log n$ for some $C_1 > 0$, on which*

$$\begin{aligned}D_n &:= \int \prod_{i=1}^n \frac{f_{\boldsymbol{\Sigma}}(\mathbf{Y}_i)}{f_{\boldsymbol{\Sigma}_{0n}}(\mathbf{Y}_i)} \mathrm{d}\Pi_n(\boldsymbol{\Sigma})\\ &\geq \exp\left(-C_2 \frac{n\eta_n^2 \log\rho_n}{\lambda^2_{p_n}(\boldsymbol{\Sigma}_{0n})}\right) \Pi_n\left(\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0n}\|_{\mathrm{F}} < \eta_n\right),\end{aligned}$$

where $C_2 > 0$ is an universal constant. If $\boldsymbol{\Sigma}_{0n} \in \mathcal{C}_{0n}$ and $\eta_n = \sqrt{s_n k_{0n}/n}$, the above inequality can be written as

$$D_n \geq \mathrm{e}^{-C_3 s_n k_{0n} \log s_n} \Pi_n\left(\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0n}\|_{\mathrm{F}} < \sqrt{s_n k_{0n}/n}\right),$$

where $C_3 > 0$ is an universal constant.

Proof. For the first part, see the proof of Lemma 9.1 of [75]. For the second assertion, we note that by (A2) and (A3), $\lambda_{p_n}(\boldsymbol{\Sigma}_{0n}) \geq c_0$ and $\lambda_1(\boldsymbol{\Sigma}_{0n})/\lambda_{p_n}(\boldsymbol{\Sigma}_{0n}) \lesssim$

$c_n \lesssim s_n$. This completes the proof. □

Lemma 3.7.5. *Let $(\epsilon_n)_{n\in\mathbb{N}}$, $(t_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ be the positive sequences. Assume that $\epsilon_n \downarrow 0$, $t_n \geq 1$ and $u_n > k_{0n}$ for any $n \in \mathbb{N}$. Assume that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{\text{iid}}{\sim} \mathsf{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Let $\mathbf{\Sigma}_{0n} := \mathbf{B}_{0n}\mathbf{B}_{0n}^\top + \sigma_{0n}^2\mathbf{I} \in \mathcal{C}_{0n}$ and $M > 2^{4/3}$. Consider the null set $H_0 := \{\mathbf{\Sigma} = \mathbf{\Sigma}_{0n}\}$ and the alternative set*

$$H_1 := \left\{\mathbf{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2\mathbf{I} : \|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| \geq M\epsilon_n, |\mathsf{supp}_{\lfloor u_n \rfloor}(\mathbf{B})| \leq t_n, \mathbf{B}_{>u_n} = \mathbf{0}\right\}.$$

Then there is a test function ϕ such that

$$\mathsf{E}_{\mathbf{\Sigma}_{0n}}\phi_n \leq 3\exp\left(2t_n \log p_n + (C_1+1)(t_n+s_n) - \frac{C_1 M^{1/2}}{(1+c_0^{-1})c_n^2} n\epsilon_n^2\right)$$

$$\sup_{\mathbf{\Sigma}:\mathbf{\Sigma}\in H_1} \mathsf{E}_{\mathbf{\Sigma}}(1-\phi_n) \leq \exp\left(C_1(t_n+s_n) - \frac{C_1 M}{4} n\epsilon_n^2\right),$$

for some universal constant $C_1 > 0$, where c_0 is in (A3).

Proof. Lemma 5.7 of [37] states that if $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{\text{iid}}{\sim} \mathsf{N}_d(\mathbf{0}, \mathbf{\Sigma})$ for some $d \times d$ positive definite matrix $\mathbf{\Sigma}$, then for any $M > 0$, there is a test function $\tilde{\phi}_n$ such that for any $M > 0$,

$$\mathsf{E}_{\mathbf{\Sigma}_1}\tilde{\phi}_n \leq \exp\left(C_1 d - \frac{C_1 M^2}{4\|\mathbf{\Sigma}_1\|^2} n\epsilon^2\right) + 2\exp(C_1 d - C_1 M^{1/2} n\epsilon^2)$$

$$\sup_{\mathbf{\Sigma}_2:\|\mathbf{\Sigma}_2-\mathbf{\Sigma}_1\|>M\epsilon} \mathsf{E}_{\mathbf{\Sigma}_2}(1-\tilde{\phi}_n) \leq \exp\left(C_1 d - \frac{C_1 M}{4}\left(1 \vee \frac{M}{(M^{1/2}+2)^2\|\mathbf{\Sigma}_1\|^2}\right) n\epsilon^2\right),$$

for some universal constant $C_1 > 0$. We decompose H_1 as

$$H_1 \subset \bigcup_{S:|S|\leq t_n} \left\{\mathbf{\Sigma} : \|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| \geq M\epsilon_n, \mathsf{supp}_{\lfloor u_n \rfloor}(\mathbf{B}) = S, \mathbf{B}_{>u_n} = \mathbf{0}\right\}.$$

Let $S := \mathsf{supp}_{\lfloor u_n \rfloor}(\mathbf{B})$, $S_0 := \mathsf{supp}_{\lfloor u_n \rfloor}(\mathbf{B}_{0n})$ and $\bar{S} := S \cup S_0$. Define

$$\mathbf{\Sigma}^{\bar{S}} := \mathbf{B}^{\bar{S}}_{\leq u_n}(\mathbf{B}^{\bar{S}}_{\leq u_n})^\top + \sigma^2\mathbf{I}, \quad \mathbf{\Sigma}^{\bar{S}}_{0n} := \mathbf{B}^{\bar{S}}_{0n,\leq u_n}(\mathbf{B}^{\bar{S}}_{0n,\leq u_n})^\top + \sigma^2_{0n}\mathbf{I}$$

where we define $\mathbf{B}^{\bar{S}}_{\leq u_n} := (\beta_{jk})_{j\in\bar{S};k\in[\lfloor u_n \rfloor]}$ and $\mathbf{B}^{\bar{S}}_{0n,\leq u_n} := (\beta_{0n,jk})_{j\in\bar{S};k\in[\lfloor u_n \rfloor]}$. Then it is easy to see that $\|\mathbf{\Sigma}^{\bar{S}} - \mathbf{\Sigma}^{\bar{S}}_{0n}\| = \|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\|$. Thus it suffices to test H_0 against

$$H_{1,S} := \left\{\mathbf{\Sigma}^{\bar{S}} : \|\mathbf{\Sigma}^{\bar{S}} - \mathbf{\Sigma}^{\bar{S}}_{0n}\| \geq M\epsilon_n\right\},$$

because $H_1 \subset \cup_{S:|S|\leq t_n} H_{1,S}$. By Lemma 5.7 of [37] with the fact that $\|\mathbf{\Sigma}^{\bar{S}}_{0n}\| = \|\mathbf{\Sigma}_{0n}\| = c_n$, there is a test ϕ^S_n satisfying for any $M > 4^{2/3}$

$$\begin{aligned}\mathsf{E}_{\mathbf{\Sigma}^{\bar{S}}_{0n}}\phi^S_n &\leq \exp\left(C_1|\bar{S}| - \frac{C_1M^2}{4c_n^2}n\epsilon_n^2\right) + 2\exp(C_1|\bar{S}| - C_1M^{1/2}n) \\ &\leq 3\exp\left(C_1|\bar{S}| - \frac{C_1M^{1/2}}{c_n^2 \vee 1}n\epsilon_n^2\right),\end{aligned}$$

and

$$\sup_{\mathbf{\Sigma}^{\bar{S}}:\mathbf{\Sigma}^{\bar{S}}\in H_{1,S}} \mathsf{E}_{\mathbf{\Sigma}^{\bar{S}}}(1-\phi^S_n) \leq \exp\left(C_1|\bar{S}| - \frac{C_1M}{4}n\epsilon_n^2\right).$$

We combine the test by $\phi_n := \max_{S:|S|\leq t_n}\phi^S_n$. Since the test function ϕ^S_n depends on the data $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ only through $(j : j \in S \cup S_0)$-th coordinates. Hence $\mathsf{E}_{\mathbf{\Sigma}_{0n}}\phi^S_n = \mathsf{E}_{\mathbf{\Sigma}^{\bar{S}}_{0n}}\phi^S_n$. Since $|\bar{S}| \leq t_n + s_n$ for any S with $|S| \leq t_n$ and

$c_n^2 \vee 1 \leq (1 + 1/c_0)c_n^2$

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\Sigma}_0}\phi_n &\leq \sum_{q=1}^{\lfloor t_n \rfloor} \binom{p_n}{q} 3\exp\left(C_1(t_n + s_n) - \frac{C_1 M^{1/2}}{c_n^2 \vee 1} n\epsilon_n^2\right) \\
&\leq 3(t_n+1)\mathrm{e}^{(t_n+1)\log p_n} \exp\left(C_1(t_n + s_n) - \frac{C_1 M^{1/2}}{(1+c_0^{-1})c_n^2} n\epsilon_n^2\right) \\
&\leq 3\exp\left((t_n+1)\log p_n + t_n + C_1(t_n + s_n) - \frac{C_1 M^{1/2}}{(1+c_0^{-1})c_n^2} n\epsilon_n^2\right) \\
&\leq 3\exp\left(2t_n \log p_n + (C_1+1)(t_n + s_n) - \frac{C_1 M^{1/2}}{(1+c_0^{-1})c_n^2} n\epsilon_n^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\sup_{\boldsymbol{\Sigma}:\boldsymbol{\Sigma}\in H_1} (1-\phi_n) &\leq \sup_{S:|S|\leq t_n} \sup_{\boldsymbol{\Sigma}^{\bar{S}}:\boldsymbol{\Sigma}^{\bar{S}}\in H_{1,S}} \mathsf{E}_{\boldsymbol{\Sigma}^{\bar{S}}}(1-\phi_n^S) \\
&\leq \exp\left(C_1(t_n + s_n) - \frac{C_1 M}{4} n\epsilon_n^2\right),
\end{aligned}$$

which completes the proof. □

Appendix A

Smooth function approximation by deep neural networks with general activation functions

A.1 Introduction

Inspired by the success of deep neural networks, many researchers have tried to give theoretical supports for the success of deep neural networks. Much of the work upto date has focused on the expressivity of deep neural networks, i.e., ability to approximate a rich class of functions efficiently. The well-known classical result on this topic is the universal approximation theorem, which states that every continuous function can be approximated arbitrarily well by a neural network [24, 45, 36, 18, 58]. However, these results do not specify the required numbers of layers and nodes of a neural network to achieve a given approximation accuracy.

Recently, several results about the effects of the numbers of layers and nodes of a deep neural network to its expressivity have been reported. They provide upper bounds of the numbers of layers and nodes required for neural networks to uniformly approximate all functions of interest. Examples of a class of functions include the space of rational functions of polynomials [91], the Hölder space [100, 80, 7, 59], Besov and mixed Besov spaces [87] and even a class of discontinuous functions [76, 46].

The nonlinear activation function is a central part that makes neural networks differ from the linear models, that is, a neural network becomes a linear function if the linear activation function is used. Therefore, the choice of an activation function substantially influences on the performance and computational efficiency. Numerous activation functions have been suggested to improve neural network learning [8, 19, 13, 78, 52, 98]. We refer to the papers [39, 78] for an overview of this topic.

As mentioned earlier, there are also many recent theoretical studies about the approximation ability of deep neural networks. However, most of the studies focus on a specific type of the activation function such as ReLU [100, 80, 76, 46, 87], or small classes of activation functions such as sigmoidal functions with additional monotonicity, continuity, and/or boundedness conditions [67, 22, 21, 23, 20] and m-admissible functions which are sufficiently smooth and bounded [7]. For definitions of sigmoidal and m-admissible functions, see [22] and [7], respectively. Thus a unified theoretical framework still lacks.

In this chapter, we investigate the approximation ability of deep neural networks with a quite general class of activation functions. We derive the required numbers of layers and nodes of a deep neural network to approximate any Hölder smooth function upto a given approximation error for the large class of activation functions. Our specified class of activation functions and the corresponding approximation ability of deep neural networks include most of previous results [100, 80, 67, 7] as special cases.

Our general theoretical results of the approximation ability of deep neural networks enables us to study statistical properties of deep neural networks. Schmidt-Hieber [80] and Chapter 1 of this thesis proved the minimax optimality of a deep neural network estimator with the ReLU activation function in regression and classification problems, respectively. In this chapter, we derive similar results for general activation functions. In addition, we apply this new result to two supervised learning problems: regression and classification.

A.1.1 Notation

We denote by $\mathbb{1}(\cdot)$ the indicator function. Let $\mathbb{R}$ be the set of real numbers and $\mathbb{N}$ be the set of natural numbers. For $m \in \mathbb{N}$, we let $[m] := \{1, \dots, m\}$. For a real valued vector $\mathbf{x} \equiv (x_1, \dots, x_d)$, we let $|\mathbf{x}|_0 := \sum_{j=1}^{d} \mathbb{1}(x_j \neq 0)$,

$|\mathbf{x}|_p := \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$ for $p \in [1,\infty)$ and $|\mathbf{x}|_\infty := \max_{1\le j\le d}|x_j|$. For simplicity, we let $|\mathbf{x}| := |\mathbf{x}|_1$. For a real valued function $f(x) : \mathbb{R} \to \mathbb{R}$, we let $f'(a), f''(a)$ and $f'''(a)$ are the first, second and third order derivatives of f at a, respectively. We let $f'(a+) := \lim_{\epsilon\downarrow 0}(f(a+\epsilon) - f(a))/\epsilon$ and $f'(a-) := \lim_{\epsilon\downarrow 0}(f(a-\epsilon) - f(a))/\epsilon$. For $x \in \mathbb{R}$, we write $(x)_+ := \max\{x, 0\}$.

A.2 Deep neural networks

In this section we provide a mathematical representation of neural networks. A neural network with $L \in \mathbb{N}$ layers, $n_l \in \mathbb{N}$ many nodes at the l-th hidden layer for $l = 1, \dots, L$, input of dimension n_0, output of dimension n_{L+1} and nonlinear activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is expressed as

$$f_\sigma(\mathbf{x}|\boldsymbol{\theta}) := \mathsf{A}_{L+1} \circ \sigma_L \circ \mathsf{A}_L \circ \cdots \circ \sigma_1 \circ \mathsf{A}_1(\mathbf{x}), \tag{A.2.1}$$

where $\mathsf{A}_l : \mathbb{R}^{n_{l-1}} \to \mathbb{R}^{n_l}$ is an affine linear map defined by $\mathsf{A}_l(\mathbf{x}) = \mathbf{W}_l\mathbf{x} + \mathbf{b}_l$ for given $n_l \times n_{l-1}$ dimensional weight matrix $\mathbf{W}_l$ and n_l dimensional bias vector $\mathbf{b}_l$ and $\sigma_l : \mathbb{R}^{n_l} \to \mathbb{R}^{n_l}$ is an element-wise nonlinear activation map defined by $\sigma_l(\mathbf{z}) := (\sigma(z_1), \dots, \sigma(z_{n_l}))^\top$. Here, $\boldsymbol{\theta}$ denotes the set of all weight matrices and bias vectors $\boldsymbol{\theta} := ((\mathbf{W}_1, \mathbf{b}_1), (\mathbf{W}_2, \mathbf{b}_2), \dots, (\mathbf{W}_{L+1}, \mathbf{b}_{L+1}))$, which we call $\boldsymbol{\theta}$ the parameter of the neural network, or simply, a network parameter.

We introduce some notations related to the network parameter. For a network parameter $\boldsymbol{\theta}$, we write $L(\boldsymbol{\theta})$ for the number of hidden layers of the corresponding neural network, and write $N_{\max}(\boldsymbol{\theta})$ for the maximum of the numbers of hidden nodes at each layer. Following a standard convention, we say that $L(\boldsymbol{\theta})$ is the depth of the neural network and $N_{\max}(\boldsymbol{\theta})$ is the width of the neural network. We let $|\boldsymbol{\theta}|_0$ be the number of nonzero elements of $\boldsymbol{\theta}$, i.e.,

$$|\boldsymbol{\theta}|_0 := \sum_{l=1}^{L+1} \left(\left|\mathrm{vec}(\mathbf{W}_l)\right|_0 + |\mathbf{b}_l|_0\right),$$

where $\mathrm{vec}(\mathbf{W}_l)$ transforms the matrix $\mathbf{W}_l$ into the corresponding vector by

concatenating the column vectors. We call $|\boldsymbol{\theta}|_0$ sparsity of the neural network. Let $|\boldsymbol{\theta}|_\infty$ be the largest absolute value of elements of $\boldsymbol{\theta}$, i.e.,

$$|\boldsymbol{\theta}|_\infty := \max\left\{\max_{1\le l\le L+1}\left|\mathrm{vec}(\mathbf{W}_l)\right|_\infty, \max_{1\le l\le L+1}|\mathbf{b}_l|_\infty\right\}.$$

We call $|\boldsymbol{\theta}|_\infty$ magnitude of the neural network. We let $\mathsf{in}(\boldsymbol{\theta})$ and $\mathsf{out}(\boldsymbol{\theta})$ be the input and output dimensions of the neural network, respectively. We denote by $\Theta_{d,o}(L,N)$ the set of network parameters with depth L, width N, input dimension d and output dimension o, that is,

$$\Theta_{d,o}(L,N) := \left\{\boldsymbol{\theta} : L(\boldsymbol{\theta}) \le L, N_{\max}(\boldsymbol{\theta}) \le N, \mathsf{in}(\boldsymbol{\theta}) = d, \mathsf{out}(\boldsymbol{\theta}) = o\right\}.$$

We further define a subset of $\Theta_{d,o}(L,N)$ with restrictions on sparsity and magnitude as

$$\Theta_{d,o}(L,N,S,B) := \left\{\boldsymbol{\theta} \in \Theta_{d,o}(L,N) : |\boldsymbol{\theta}|_0 \le S, |\boldsymbol{\theta}|_\infty \le B\right\}.$$

A.3 Classes of activation functions

In this section, we consider two classes of activation functions. These two classes include most of commonly used activation functions. Definitions and examples of each class of activation functions are provided in the consecutive two subsections.

A.3.1 Piecewise linear activation functions

We first consider piecewise linear activation functions.

Definition A.3.1. A function $\sigma : \mathbb{R} \to \mathbb{R}$ is continuous piecewise linear if it is continuous and there exist a finite number of break points $a_1 \le a_2 \le \cdots \le a_K \in \mathbb{R}$ with $K \in \mathbb{N}$ such that $\sigma'(a_k-) \ne \sigma'(a_k+)$ for every $k = 1, \ldots, K$ and $\sigma(x)$ is linear on $(-\infty, a_1], [a_1, a_2], \ldots, [a_{K-1}, a_K], [a_K, \infty)$.

Throughout this paper, we write "picewise linear" instead of "continuous picewise linear" for notational simplicity unless there is a confusion. The representative examples of piecewise linear activation functions are as follows:

- ReLU: $\sigma(x) = \max\{x, 0\}$.
- Leaky ReLU: : $\sigma(x) = \max\{x, ax\}$ for $a \in (0, 1)$.

The ReLU activation function is the most popular choice in practical applications due to better gradient propagation and efficient computation [39]. In this reason, most of the recent results on the function approximation by deep neural networks are based on the ReLU activation function [100, 80, 76, 46, 87]. In Section A.4, as Yarotsky [100] did, we extend these results to any continuous piecewise linear activation function by showing that the ReLU activation function can be exactly represented by a linear combination of piecewise linear activation functions. A formal proof for this argument is presented in Section A.6.1.

A.3.2 Locally quadratic activation functions

One of the basic building blocks in approximation by neural networks is the square function, which should be approximated precisely. Piecewise linear activation functions have zero curvature (i.e., constant first-order derivative) inside each interval divided by its break points, which makes it relatively difficult to approximate the square function efficiently. But if there is an interval on which the activation function has nonzero curvature, the square function can be approximated more efficiently, which is a main motivation of considering a new class of activation functions that we call locally quadratic.

Definition A.3.2. A function $\sigma : \mathbb{R} \to \mathbb{R}$ is locally quadratic if there exits an open interval $(a, b) \subset \mathbb{R}$ on which σ is three times continuously differentiable with bounded derivatives and there exists $t \in (a, b)$ such that $\sigma'(t) \neq 0$ and $\sigma''(t) \neq 0$.

We now give examples of locally quadratic activation functions. First of all, any nonlinear smooth activation function with nonzero second derivative, is locally quadratic. Examples are:

- Sigmoid: $\sigma(x) = \dfrac{1}{1+\mathrm{e}^{-x}}$.
- Tangent hyperbolic: $\sigma(x) = \dfrac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}}$.
- Inverse square root unit (ISRU) [13]: $\sigma(x) = \dfrac{x}{\sqrt{1+ax^2}}$ for $a > 0$.
- Soft clipping [52]: $\sigma(x) = \dfrac{1}{a}\log\left(\dfrac{1+\mathrm{e}^{ax}}{1+\mathrm{e}^{a(x-1)}}\right)$ for $a > 0$.
- SoftPlus [39]: $\sigma(x) = \log(1+\mathrm{e}^{x})$.
- Swish [78]: $\sigma(x) = \dfrac{x}{1+\mathrm{e}^{-x}}$.

In addition, piecewise smooth function having nonzero second derivative on at least one piece, is also locally quadratic. Examples are:

- Rectified power unit (RePU) [59]: $\sigma(x) = \max\{x^k, 0\}$ for $k \in \mathbb{N} \setminus \{1\}$.
- Exponential linear unit (ELU) [19]: $\sigma(x) = a(\mathrm{e}^{x}-1)\mathbb{1}(x \le 0) + x\mathbb{1}(x > 0)$ for $a > 0$. : $\sigma(x) = \dfrac{x}{\sqrt{1+ax^2}}\mathbb{1}(x \le 0) + x\mathbb{1}(x > 0)$ for $a > 0$.
- Softsign [8]: $\sigma(x) = \dfrac{x}{1+|x|}$.
- Square nonlinearity [98]:
  $\sigma(x) = \mathbb{1}(x > 2) + (x - x^2/4)\mathbb{1}(0 \le x \le 2) + (x + x^2/4)\mathbb{1}(-2 \le x < 0) - \mathbb{1}(x < -2)$.

A.4 Approximation of Hölder smooth functions by deep neural networks

In this section we introduce the function class we consider and show the approximation ability of the deep neural networks with an activation function considered in Section A.3.

We recall the definition of Hölder smooth functions. For a d-dimensional multiple index $\mathbf{m} \equiv (m_1, \ldots, m_d) \in \mathbb{N}_0^d$ where $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, we let $\mathbf{x}^{\mathbf{m}} := x_1^{m_1} \cdots x_d^{m_d}$ for $\mathbf{x} \in \mathbb{R}^d$. For a function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ denotes the domain of the function, we let $\|f\|_\infty := \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})|$. We use notation

$$\partial^{\mathbf{m}} f := \frac{\partial^{|\mathbf{m}|} f}{\partial \mathbf{x}^{\mathbf{m}}} = \frac{\partial^{|\mathbf{m}|} f}{\partial x_1^{m_1} \cdots \partial x_d^{m_d}},$$

for $\mathbf{m} \in \mathbb{N}_0^d$ to denote a derivative of f of order $\mathbf{m}$. We denote by $\mathcal{C}^m(\mathcal{X})$, the space of m times differentiable functions on $\mathcal{X}$ whose partial derivatives of order $\mathbf{m}$ with $|\mathbf{m}| \leq m$ are continuous. We define the Hölder coefficient of order $s \in (0, 1]$ as

$$[f]_s := \sup_{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, \mathbf{x}_1 \neq \mathbf{x}_2} \frac{|f(\mathbf{x}_1) - f(\mathbf{x}_2)|}{|\mathbf{x}_1 - \mathbf{x}_2|^s}.$$

For a positive real value α, the Hölder space of order α is defined as

$$\mathcal{H}^\alpha(\mathcal{X}) := \left\{ f \in \mathcal{C}^{\lfloor \alpha \rfloor}(\mathcal{X}) : \|f\|_{\mathcal{H}^\alpha(\mathcal{X})} < \infty \right\},$$

where $\|f\|_{\mathcal{H}^\alpha(\mathcal{X})}$ denotes the Hölder norm defined by

$$\|f\|_{\mathcal{H}^\alpha(\mathcal{X})} := \sum_{\mathbf{m} \in \mathbb{N}_0^d : |\mathbf{m}| \leq \lfloor \alpha \rfloor} \|\partial^{\mathbf{m}} f\|_\infty + \sum_{\mathbf{m} \in \mathbb{N}_0^d : |\mathbf{m}| = \lfloor \alpha \rfloor} [\partial^{\mathbf{m}} f]_{\alpha - \lfloor \alpha \rfloor}.$$

We denote by $\mathcal{H}^{\alpha,R}(\mathcal{X})$ the closed ball in the Hölder space of radius R with respect to the Hölder norm, i.e.,

$$\mathcal{H}^{\alpha,R}(\mathcal{X}) := \left\{ f \in \mathcal{H}^{\alpha}(\mathcal{X}) : \|f\|_{\mathcal{H}^{\alpha}(\mathcal{X})} \leq R \right\}.$$

We now ready to present our main theorem.

Theorem A.4.1. *Let $d \in \mathbb{N}$, $\alpha > 0$ and $R > 0$. Let the activation function σ be either continuous piecewise linear or locally quadratic. Let $f \in \mathcal{H}^{\alpha,R}([0,1]^d)$. Then there exist positive constants L_0, N_0, S_0 and B_0 depending only on d, α, R and σ such that, for any $\epsilon > 0$, there is a neural network*

$$\boldsymbol{\theta}_\epsilon \in \Theta_{d,1}\left(L_0 \log(1/\epsilon), N_0 \epsilon^{-d/\alpha}, S_0 \epsilon^{-d/\alpha} \log(1/\epsilon), B_0 \epsilon^{-4(d/\alpha+1)}\right) \quad \text{(A.4.1)}$$

satisfying

$$\sup_{\mathbf{x}\in[0,1]^d} \left| f(\mathbf{x}) - f_\sigma(\mathbf{x}|\boldsymbol{\theta}_\epsilon) \right| \leq \epsilon. \quad \text{(A.4.2)}$$

Proof. The proof is different by the two classes of activation functions. For piecewise linear activation functions see Section A.6.1 and for locally quadratic activation functions, see Section A.6.2. □

The result of Theorem A.4.1 is equivalent to the results on the approximation by ReLU neural networks [100, 80] in a sense that the upper bounds of the depth, width and sparsity are the same orders of those for ReLU, namely, depth $= O(\log(\epsilon^{-1}))$, width $= O(\epsilon^{-d/\alpha})$ and sparsity $= O(\epsilon^{-d/\alpha}\log(\epsilon^{-1}))$. We remark that each upper bound is equivalent to the corresponding lower bound established by [100] up to logarithmic factor.

For piecewise linear activation functions, Yarotsky [100] derived similar results to ours. For locally quadratic activation functions, only special classes of activation functions were considered in the previous work. Li et al. [59] considered the RePU activation function and Bauer and Kohler [7] considered sufficiently smooth and bounded activation functions which include the sigmoid, tangent hyperbolic, ISRU and soft clipping activation functions. However, soft plus, swish, ELU, ISRLU, softsign and square nonlinearity activation functions are new ones only considered in our results.

Even if the orders of the depth, width and sparsity are the same for both both piecewise linear and locally quadratic activation functions, the ways of approximating a smooth function by use of these two activation function classes are quite different. To describe this point, let us provide an outline of the proof. We first consider equally spaced grid points with length $1/M$ inside the d-dimensional unit hypercube $[0,1]^d$. Let $\mathbb{G}_{d,M}$ be the set of such grid points, namely,

$$\mathbb{G}_{d,M} := \left\{ \frac{1}{M}(m_1, \ldots, m_d) : m_j \in \{0, 1, \ldots, M\}, j = 1, \ldots, d \right\}.$$

For a given Hölder smooth function f of order α, we first find a "local" function for each grid that approximates the target function near the grid point but vanishes at apart from the grid point. To be more specific, we construct the local functions $g_{\mathbf{z}}$, $\mathbf{z} \in \mathbb{G}_{d,M}$ which satisfies:

$$\sup_{x \in [0,1]^d} \left| f(\mathbf{x}) - \sum_{\mathbf{z} \in \mathbb{G}_{d,M}} g_{\mathbf{z},M}(\mathbf{x}) \right| \leq C |\mathbb{G}_{d,M}|^{-\alpha/d}, \tag{A.4.3}$$

for some universal constant $C > 0$. The inequality (A.4.3) implies that the more grid points we used, the more accurate approximation we get. Moreover, the quality of approximation is improved when the target function is more smooth (i.e., large α) and low dimensional (i.e., small d). In fact, $g_{\mathbf{z},M}(\mathbf{x})$ is given by a product of the Taylor polynomial

$$P_{\mathbf{z},M}(\mathbf{x}) := \sum_{\mathbf{m} \in \mathbb{N}_0^d : |\mathbf{m}| \leq \alpha} \left(\partial^{\mathbf{m}} f\right)(\mathbf{z}) \frac{(\mathbf{x} - \mathbf{z})^{\mathbf{m}}}{\mathbf{m}!}$$

at $\mathbf{z}$ and the local basis function

$$\phi_{\mathbf{z},M}(\mathbf{x}) := \prod_{j=1}^{d} (1/M - |x_j - z_j|)_+,$$

where $\mathbf{m}! := \prod_{j=1}^{d} m_j!$. By simple algebra, we have

$$P_M(\mathbf{x}) := \sum_{\mathbf{z}\in\mathbb{G}_{d,M}} g_{\mathbf{z},M}(\mathbf{x}) := \sum_{\mathbf{z}\in\mathbb{G}_{d,M}} P_{\mathbf{z},M}(\mathbf{x})\phi_{\mathbf{z},M}(\mathbf{x})$$
$$= \sum_{\mathbf{z}\in\mathbb{G}_{d,M}} \sum_{\mathbf{m}:|\mathbf{m}|\le\alpha} \beta_{\mathbf{z},\mathbf{m}}\mathbf{x}^{\mathbf{m}}\phi_{\mathbf{z},M}(\mathbf{x}),$$

where $\beta_{\mathbf{z},\mathbf{m}} := \sum_{\tilde{\mathbf{m}}:\tilde{\mathbf{m}}\ge\mathbf{m},|\tilde{\mathbf{m}}|\le\alpha} \left(\partial^{\tilde{\mathbf{m}}} f\right)(\mathbf{z}) \frac{(-\mathbf{z})^{\tilde{\mathbf{m}}-\mathbf{m}}}{\mathbf{m}!(\tilde{\mathbf{m}}-\mathbf{m})!}$.

The second stage is to approximate each monomial $\mathbf{x}^{\mathbf{m}}$ and each local basis function $\phi_{\mathbf{z},M}(\mathbf{x})$ by a neural network. Each monomial can be approximated more efficiently by a neural network with a locally quadratic activation function than a piecewise linear activation function since each monomial has nonzero curvature. On the other hand, the local basis function can be approximated more efficiently by a neural network with a piecewise linear activation than a locally quadratic activation function since the local basis function is piecewise linear itself. That is, there is a trade-off in using either a piecewise linear or a locally quadratic activation function.

We close this section by giving a comparison of our result to the approximation error analysis of [7]. Bauer and Kohler [7] studies approximation of the Hölder smooth function of order α by a two layer neural network with m-admissible activation functions with $m \ge \alpha$, where a function σ is called m-admissible if (1) σ is at least $m+1$ times continuously differentiable with bounded derivatives; (2) a point $t \in \mathbb{R}$ exists, where all derivatives up to the order m of σ are different from zero; and (3) $|\sigma(x) - 1| \le 1/x$ for $x > 0$ and $|\sigma(x)| \le 1/|x|$ for $x < 0$. Our notion of locally quadratic activation functions is a generalized version of the m-admissibility.

In the proof of [7], the condition $m \ge \alpha$ is necessary because they approximate any monomial of order $\mathbf{m}$ with $|\mathbf{m}| \le \alpha$ with a two layer neural network, which is impossible when $m < \alpha$. We drop the condition $m \ge \alpha$ by showing that any monomial of order $\mathbf{m}$ with $|\mathbf{m}| \le \alpha$ can be approximated by a neural network with a finite number of layers, which depends on α.

A.5 Application to statistical learning theory

In this section, we apply our results about the approximation error of neural networks to the supervised learning problems of regression and classification. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ the output space. Let $\mathcal{F}$ be a given class of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. Let P be the true but unknown data generating distribution on $\mathcal{X} \times \mathcal{Y}$. The aim of supervised learning is to find a predictive function that minimizes the population risk $\mathcal{R}(f) := \mathsf{E}\ell(Y, f(\mathbf{X}))$ with respect to a given loss function ℓ. Since P_0 is unknown, we cannot directly minimize the population risk, and thus any estimator $\hat{f}$ inevitably has the excess risk which is defined as $\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f)$. For a given sample of size n, let $\mathcal{F}_n$ be a given subset of $\mathcal{F}$ called a sieve and let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be observed (training) data of input–output pairs assumed to be independent realizations of $(\mathbf{X}, Y)$ following P. Let $\hat{f}_n$ be an estimated function among $\mathcal{F}_n$ based on the training data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. The excess risk of $\hat{f}_n$ is decomposed to approximation and estimation errors as

$$\mathcal{R}(\hat{f}_n) - \inf_{f \in \mathcal{F}} \mathcal{R}(f) = \underbrace{\left[\mathcal{R}(\hat{f}_n) - \inf_{f \in \mathcal{F}_n} \mathcal{R}(f)\right]}_{\text{Estimation error}} + \underbrace{\left[\inf_{f \in \mathcal{F}_n} \mathcal{R}(f) - \inf_{f \in \mathcal{F}} \mathcal{R}(f)\right]}_{\text{Approximation error}}. \tag{A.5.1}$$

There is a trade-off between approximation and estimation errors. If the function class $\mathcal{F}_n$ is sufficiently complex to approximate the optimal estimator given by

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$$

well, then the estimation error becomes large due to high variance. In contrast, if $\mathcal{F}_n$ is small, it leads to low estimation error but it suffers from large approximation error.

One of the advantages of neural networks is that we can construct a sieve which has good approximation ability as well as low complexity. Schmidt-Hieber [80] and Chapter 1 of this thesis proved that a neural network estimator can achieve the optimal balance between the approximation and estimation errors to obtain the minimax optimal convergence rates in regression and classification problems, respectively. But they only considered the ReLU activation function. Based on the results of Theorem A.4.1, we can

easily extend their results to general activation functions.

The main tool to derive the minimax optimal convergence rate is that the complexity of a class of functions generated by a neural network is not affected much by a choice of an activation function, provided that the activation function is Lipschitz continuous. The function $\sigma : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous if there is a constant $C_\sigma > 0$ such that

$$|\sigma(x_1) - \sigma(x_2)| \leq C_\sigma |x_1 - x_2|, \tag{A.5.2}$$

for any $x_1, x_2 \in \mathbb{R}$. Here, C_σ is called the Lipschitz constant. We use the covering number with respect to the L_∞ norm $\|\cdot\|_\infty$ as a measure of complexity of function classes. We recall the definition of the covering number. For the definition of the covering number, see Section 1.7.1. The following proposition provides the covering number of a class of functions generated by neural networks.

Proposition A.5.1. *Assume that the activation function σ is Lipschitz continuous with the Lipschitz constant C_σ. Consider a class of functions generated by a neural network*

$$\mathcal{F}_{d,1}(L, N, S, B) := \left\{ f_\sigma(\cdot|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{d,1}(L, N, S, B) \right\}.$$

Then for any $\delta > 0$,

$$\begin{aligned} &\log \mathcal{N}\left(\delta, \mathcal{F}_{d,1}(L, N, S, B), \|\cdot\|_\infty\right) \\ &\leq 2L(S+1) \log\left(\delta^{-1} C_\sigma L (N+1)(B \vee 1)\right), \end{aligned}$$

where $B \vee 1 := \max\{B, 1\}$.

Proof. See Section A.6.3. □

The result in Proposition A.5.1 is very similar to the existing results in literature, e.g., Theorem 14.5 of [2], Lemma 5 of [80] and Lemma 3 of [87]. We employ similar techniques used in [2, 80, 87] to obtain the version presented here. We give the proof of this proposition in Section A.6.3.

All of the activation functions considered in Section A.3 except RePU satisfy the Lipschitz (A.5.2) and hence Proposition A.5.1 can be applied. An interesting implication of Proposition A.5.1 is that the complexity of the function class generated by neural networks is not affected by the choice of

an activation function. Hence, the remaining step to derive the convergence rate of a neural network estimator is that approximation accuracies by various activation functions are the same as that of the ReLU neural network.

A.5.1 Application to regression

First we consider the regression problem. For simplicity, we let $\mathcal{X} = [0,1]^d$. Suppose that the generated model is $Y|\mathbf{X} = \mathbf{x} \sim \mathsf{N}(f_0(\mathbf{x}), 1)$ for some $f_0 : [0,1]^d \to \mathbb{R}$, and $\mathbf{X} \sim \mathsf{P}_\mathbf{X}$. The performance of an estimates f is measured by the $L_2(\mathsf{P}_\mathbf{X})$ distance to the true function f_0, i.e.,

$$\|f - f_0\|_{2,\mathsf{P}_\mathbf{X}} := \left(\int (f(\mathbf{x}) - f_0(\mathbf{x})) \mathrm{d}\mathsf{P}_\mathbf{X}(\mathbf{x}) \right)^{1/2}.$$

The following theorem proves that the optimal convergence rate is obtained by the deep neural network estimator of the regression function f_0 for a general activation function.

Theorem A.5.2. *Suppose that the activation function σ is either piecewise linear or locally quadratic satisfying the Lipschitz condition (A.5.2). Then there are universal positive constants L_0, N_0, S_0 and B_0 such that the deep neural network estimator obtained by*

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}_{\sigma,n}} \sum_{i=1}^{n} \left(Y_i - f(\mathbf{X}_i) \right)^2,$$

with

$$\mathcal{F}_{\sigma,n} := \Big\{ f_\sigma(\cdot|\boldsymbol{\theta}) : \big\| f_\sigma(\cdot|\boldsymbol{\theta}) \big\|_\infty \le 2R, \\ \boldsymbol{\theta} \in \Theta_{d,1} \left(L_0 \log n, N_0 n^{\frac{d}{2\alpha+d}}, S_0 n^{\frac{d}{2\alpha+d}} \log n, B_0 n^\kappa \right) \Big\}$$

for some $\kappa > 0$ satisfies

$$\sup_{\mathsf{P}: f_0 \in \mathcal{H}^{\alpha,R}([0,1]^d)} \mathsf{E} \left[\|\hat{f}_n - f_0\|_{2,\mathsf{P}_\mathbf{X}}^2 \right] \lesssim n^{-\frac{2\alpha}{2\alpha+d}} \log^3 n,$$

where the expectation is taken over the training data.

Proof. See Section A.6.4. □

A.5.2 Application to binary classification

The aim of the binary classification is to find a classifier that predicts the label $y \in \{-1, 1\}$ for any input $\mathbf{x} \in [0,1]^d$. An usual assumption on the data generating process is that $Y|\mathbf{X} = \mathbf{x} \sim 2\text{Bernoulli}(\eta(\mathbf{x})) - 1$ for some $\eta : [0,1]^d \to [0,1]$, where $\text{Bernoulli}(p)$ denotes the Bernoulli distribution with parameter p. Note that $\eta(\mathbf{x})$ is the conditional probability function $\mathsf{P}_0(Y = 1|\mathbf{X} = \mathbf{x})$. A common approach is, instead of finding a classifier directly, to construct a real valued function f, a so-called classification function, and predict the label y based on the sign of $f(\mathbf{x})$. The performance of a classification function f is measured by the misclassification risk $\mathcal{R}_{01}(f)$, which is defined by

$$\mathcal{R}_{01}(f) := \mathsf{E}\mathbb{1}(Yf(\mathbf{X}) < 0).$$

Let f^* be the Bayes classifier defined by

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{R}_{01}(f)$$

where $\mathcal{F}$ is the class of all real-valued measurable functions on $[0,1]^d$.

It is well known that the convergence rate of the excess risk for classification is faster than that of regression when the conditional probability function $\eta(\mathbf{x})$ satisfies the following condition: there is a constant $q \in [0, \infty]$ such that for any sufficiently small $u > 0$, we have

$$\mathsf{P}_{\mathbf{X}}\left(|\eta(\mathbf{X}) - 1/2| < u\right) \leq u^q. \tag{A.5.3}$$

This condition is called the Tsybakov noise condition and q is called the noise exponent [63, 93]. When q is larger, the classification task is easier since the probability of generating vague samples become smaller. The following theorem proves that the optimal convergence rate can be obtained by the deep neural network estimator with an activation function considered in

Section A.3. As is done by Chapter 1, we consider the hinge loss $\ell_{\mathsf{hinge}}(z) := \max\{1-z, 0\}$.

Theorem A.5.3. *Let $\mathcal{P}_q$ be a distribution on $[0,1]^d \times \{-1,1\}$ satisfying the Tsybakov noise condition (A.5.3) with the noise exponent $q \in [0,\infty]$. Suppose that the activation function σ, which is either piecewise linear or locally quadratic satisfying the Lipschitz condition (A.5.2), is used for all hidden layers except the last one and the ReLU activation function is used for the last hidden layer. Then there are universal positive constants L_0, N_0, S_0 and B_0 such that the deep neural network estimator obtained by*

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}_{\sigma,n}} \sum_{i=1}^{n} \ell_{\mathsf{hinge}}(Y_i f(\mathbf{X}_i)),$$

with

$$\mathcal{F}_{\sigma,n} := \Big\{ f_\sigma(\cdot|\boldsymbol{\theta}) : \big\|f_\sigma(\cdot|\boldsymbol{\theta})\big\|_\infty \le 1, \\ \boldsymbol{\theta} \in \Theta_{d,1}\left(L_0 \log n, N_0 n^\nu \log^{-3\nu} n, S_0 n^\nu \log^{-3\nu+1} n, B_0 n^\kappa\right) \Big\},$$

for $\nu := d/\{\alpha(q+2)+d\}$ and some $\kappa > 0$ satisfies

$$\sup_{\mathsf{P}:\eta \in \mathcal{H}^{\alpha,R}([0,1]^d)} \mathsf{E}\left[\mathcal{R}_{01}(\hat{f}_n) - \mathcal{R}_{01}(f^*)\right] \lesssim \left(\frac{\log^3 n}{n}\right)^{\frac{\alpha(q+1)}{\alpha(q+2)+d}},$$

where the expectation is taken over the training data.

Proof. See Section A.6.5. □

Note that the Bayes classifier $f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_{01}(f)$ is given by

$$f^*(\mathbf{x}) = 2\mathbb{1}\left(2\eta(\mathbf{x}) - 1 \ge 0\right) - 1,$$

which is an indicator function. Since a neural network with the ReLU activation function can approximate indicator functions well [76, 46, 51], we use the ReLU activation function in the last layer in order to approximate the

Bayes classifier more precisely and thus to achieve the optimal convergence rate.

A.6 Proofs

A.6.1 Proof of Theorem A.4.1 for piecewise linear activation functions

The main idea of the proof is that any deep neural network with the ReLU activation function can be exactly reconstructed by a neural network with a piecewise activation function whose proof is in the next lemma that is a slight modification of Proposition 1 (b) of [100].

Lemma A.6.1. *Let σ be an any continuous peicewise linear activation function, and ρ be the ReLU activation function. Let $\boldsymbol{\theta} \in \Theta_{d,1}(L, N, S, B)$. Then there exists*

$$\boldsymbol{\theta}^* \in \Theta_{d,1}(L, 2N, 4S + 2LN + 1, C_1 B)$$

such that

$$\sup_{\mathbf{x}\in[0,1]^d} \left| f_\sigma(\mathbf{x}|\boldsymbol{\theta}^*) - f_\rho(\mathbf{x}|\boldsymbol{\theta}) \right| = 0,$$

where $C_1 > 0$ is a constant depending on the activation function σ.

Proof. Let a be any break point of σ. Note that $\sigma(a-) \neq \sigma(a+)$. Let r_0 be the distance between a and the closest other break point. Then σ is linear on $[a - r_0, a]$ and $[a, a + r_0]$. Then for any $r > 0$, the ReLU activation function $\rho(x) := (x)_+$ is expressed as

$$\begin{aligned} \rho(x) &= \frac{\sigma\left(a + \frac{r_0}{2r}x\right) - \sigma\left(a - \frac{r_0}{2} + \frac{r_0}{2r}x\right) - \sigma(a) + \sigma\left(a - \frac{r_0}{2}\right)}{\left(\sigma'(a+) - \sigma'(a-)\right)\frac{r_0}{2r}} \\ &=: u_1\sigma\left(a + \frac{r_0}{2r}x\right) + u_2\sigma\left(a - \frac{r_0}{2} + \frac{r_0}{2r}x\right) + v \end{aligned} \tag{A.6.1}$$

for any $x \in [-r, r]$, where we define

$$u_1 := 1/\left(\left(\sigma'(a+) - \sigma'(a-)\right)\frac{r_0}{2r}\right)$$
$$u_2 := -1/\left(\left(\sigma'(a+) - \sigma'(a-)\right)\frac{r_0}{2r}\right)$$
$$v := \left(-\sigma(a) + \sigma\left(a - r_0/2\right)\right)/\left(\left(\sigma'(a+) - \sigma'(a-)\right)\frac{r_0}{2r}\right).$$

Let $\boldsymbol{\theta} \equiv ((\mathbf{W}_1, \mathbf{b}_1), \ldots, (\mathbf{W}_{L+1}, \mathbf{b}_{L+1})) \in \Theta_{d,1}(L, N, S, B)$ be given. Since both input $\mathbf{x} \in [0,1]^d$ and the network parameter $\boldsymbol{\theta}$ are bounded, we can take a sufficiently large r so that Equation (A.6.1) holds for any hidden nodes of the network $\boldsymbol{\theta}$.

We consider the neural network $\boldsymbol{\theta}^* \equiv ((\mathbf{W}_1^*, \mathbf{b}_1^*), \ldots, (\mathbf{W}_{L+1}^*, \mathbf{b}_{L+1}^*)) \in \Theta_{d,1}(L, 2N)$, where we set

$$\mathbf{W}_l^* := \frac{r_0}{2r}\begin{pmatrix} u_1\mathbf{W}_l & u_2\mathbf{W}_l \\ u_1\mathbf{W}_l & u_2\mathbf{W}_l \end{pmatrix} \in \mathbb{R}^{2n_l \times 2n_{l-1}},$$
$$\mathbf{b}_l^* := \begin{pmatrix} a\mathbf{1}_{n_l} + \frac{r_0}{2r}(v\mathbf{W}_l\mathbf{1}_{n_{l-1}} + \mathbf{b}_l) \\ \left(a - \frac{r_0}{2}\right)\mathbf{1}_{n_l} + + \frac{r_0}{2r}(v\mathbf{W}_l\mathbf{1}_{n_{l-1}} + \mathbf{b}_l) \end{pmatrix} \in \mathbb{R}^{2n_l},$$

for $l = 1, \ldots, L$ and

$$\mathbf{W}_{L+1}^* := \begin{pmatrix} u_1\mathbf{W}_{L+1} & u_2\mathbf{W}_{L+1} \end{pmatrix}, \quad \mathbf{b}_{L+1}^* := v.$$

Here, $\mathbf{1}_n$ denotes the n-dimensional vector of $1's$. Then by Equation (A.6.1) and some algebra, we have that $f_\sigma(\mathbf{x}|\boldsymbol{\theta}^*) = f_\rho(\mathbf{x}|\boldsymbol{\theta})$ for any $\mathbf{x} \in [0,1]^d$. For the sparsity of $\boldsymbol{\theta}^*$, we note that

$$\left|\mathrm{vec}(\mathbf{W}_l^*)\right|_0 + \left|\mathbf{b}_l^*\right|_0 \le 4\left|\mathrm{vec}(\mathbf{W}_l)\right|_0 + 2N_{\max}(\boldsymbol{\theta})$$

which implies that $|\boldsymbol{\theta}^*|_0 \le 4|\boldsymbol{\theta}|_0 + 2L(\boldsymbol{\theta})N_{\max}(\boldsymbol{\theta}) + 1$. □

Thanks to Lemma A.6.1, to prove Theorem A.4.1 for piecewise linear activation functions, it suffices to show the approximation ability of the ReLU networks, which is already done by [80] as in the next lemma.

Lemma A.6.2 (Theorem 5 of [80]). *Let ρ be the ReLU activation function. For any $f \in \mathcal{H}^{\alpha,R}([0,1]^d)$ and any integers $m \ge 1$ and $M \ge \max\left\{(\alpha+1)^d, (R+1)\mathrm{e}^d\right\}$,*

there exists a network parameter $\boldsymbol{\theta} \in \Theta_{d,1}(L, N, S, 1)$ *such that*

$$\sup_{\mathbf{x}\in[0,1]^d} \left| f_\rho(\mathbf{x}|\boldsymbol{\theta}) - f(\mathbf{x}) \right| \leq (2R+1)(1+d^2+\alpha^2)6^d M 2^{-m} + R3^\alpha M^{-\alpha/d}, \tag{A.6.2}$$

where $L = 8 + (m+5)(1 + \lceil \log_2(d \vee \alpha) \rceil)$, $N = 6(d + \lceil \alpha \rceil)M$, *and* $S = 141(d+\alpha+1)^{3+d}M(m+6)$.

Theorem A.4.1 for piecewise linear activation functions is a direct consequence of Lemma A.6.1 and Lemma A.6.2, which is summarized as follows.

Proof of Theorem A.4.1 for piecewise linear activation functions. Let ρ be the ReLU activation function. By letting $M = 3^d(2R)^{d/\alpha}\epsilon^{-d/\alpha}$ and

$$m = \log_2 \left(2(2R+1)(1+d^2+\alpha^2)18^d(2R)^{d/\alpha}\epsilon^{-d/\alpha-1} \right),$$

Lemma A.6.2 implies that there exists a network parameter $\boldsymbol{\theta}'$ such that

$$\sup_{\mathbf{x}\in[0,1]^d} |f_\rho(\mathbf{x}|\boldsymbol{\theta}') - f(\mathbf{x})| \leq \epsilon$$

with $L(\boldsymbol{\theta}') \leq L_0' \log(1/\epsilon)$, $N_{\max}(\boldsymbol{\theta}') \leq N_0'\epsilon^{-d/\alpha}$ and $|\boldsymbol{\theta}'|_0 \leq S_0'\epsilon^{-d/\alpha}\log(1/\epsilon)$ for some positive constants L_0', N_0', and S_0' depending only on α, d and R. Hence by Lemma A.6.1, there is a network parameter $\boldsymbol{\theta}$ producing the same output of the ReLU neural network $f_\rho(\cdot|\boldsymbol{\theta})$ with $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}')$, $N_{\max}(\boldsymbol{\theta}) = 2N_{\max}(\boldsymbol{\theta}')$, $|\boldsymbol{\theta}|_0 \leq 4|\boldsymbol{\theta}'|_0 + 2L(\boldsymbol{\theta}')N_{\max}(\boldsymbol{\theta}') + 1 \leq S_0\epsilon^{-d/\alpha}\log(1/\epsilon)$ and $|\boldsymbol{\theta}|_\infty \leq B_0|\boldsymbol{\theta}'|_\infty$ for some $S_0 > 0$ depending only on α, d, R and σ, and some $B_0 > 0$ depending only on σ, which completes the proof. □

A.6.2 Proof of Theorem A.4.1 for locally quadratic activation functions

Lemma A.6.3. *Assume that an activation function* σ *is locally quadratic. There is a constant* K_0 *depending only on the activation function such that for any* $K > K_0$ *the following results hold.*

(a) *There is a neural network* $\boldsymbol{\theta}_2 \in \Theta_{1,1}(1,3)$ *with* $|\boldsymbol{\theta}_2|_\infty \leq K^2$ *such that*

$$\sup_{x\in[-1,1]} \left| f_\sigma(x|\boldsymbol{\theta}_2) - x^2 \right| \leq \frac{C_1}{K},$$

where $C_1 > 0$ *is a constant depending only on* σ.

(b) *Let* $A > 0$. *There is a neural network parameter* $\boldsymbol{\theta}_{\times,A} \in \Theta_{2,1}(1,9)$ *with* $\left|\boldsymbol{\theta}_{\times,A}\right|_\infty \leq \max\{K^2, 2A^2\}$ *such that*

$$\sup_{\mathbf{x}\in[-A,A]^2} \left| f_\sigma(\mathbf{x}|\boldsymbol{\theta}_{\times,A}) - x_1 x_2 \right| \leq \frac{6A^2 C_1}{K}.$$

(c) *Let* α *be a positive integer. For any multi-index* $\mathbf{m} \in \mathbb{N}_0^d$ *with* $|\mathbf{m}| \leq \alpha$, *there is a network parameter* $\boldsymbol{\theta}_{\mathbf{m}} \in \Theta_{d,1}(\lceil \log_2 \alpha \rceil, 9\alpha)$ *with* $|\boldsymbol{\theta}_{\mathbf{m}}|_\infty \leq \max\{K^2, C_2\}$ *such that*

$$\sup_{\mathbf{x}\in[0,1]^d} \left| f_\sigma(\mathbf{x}|\boldsymbol{\theta}_{\mathbf{m}}) - \mathbf{x}^{\mathbf{m}} \right| \leq \frac{C_3}{K},$$

for some positive constants C_2 *and* C_3 *depending only on* σ *and* α.

(d) *There is a network parameter* $\boldsymbol{\theta}_{1/2} \in \Theta_{1,1}(\lceil \log K \rceil, 15)$ *with* $\left|\boldsymbol{\theta}_{1/2}\right|_\infty \leq \max\{K^2, C_4\}$ *such that*

$$\sup_{x\in[0,2]} \left| f_\sigma(x|\boldsymbol{\theta}_{1/2}) - \sqrt{x} \right| \leq C_5 \frac{\log K}{K}$$

for some positive constants C_4 *and* C_5 *depending only on* σ.

(e) *There is a network parameter* $\boldsymbol{\theta}_{\text{abs}} \in \Theta_{1,1}(\lceil \log K \rceil, 15)$ *with* $|\boldsymbol{\theta}_{\text{abs}}|_\infty \leq \max\{K^2, C_6\}$ *such that*

$$\sup_{x\in[-1,1]} \left| f_\sigma(x|\boldsymbol{\theta}_{\text{abs}}) - |x| \right| \leq \frac{C_7}{\sqrt{K}},$$

for some positive constants C_6 *and* C_7 *depending only on* σ.

Proof. Recall that there is an interval (a,b) on which $\sigma(x)$ is three times continuously differentiable with bounded derivatives and there is $t \in (a,b)$ such that $\sigma'(t) \neq 0$ and $\sigma''(t) \neq 0$

Proof of (a). Take K large so that $2/K < \min\{|t-b|, |t-a|\}$. Consider a neural network

$$f_\sigma(x|\boldsymbol{\theta}_2) := \sum_{k=0}^{2} (-1)^{k-1} \frac{K^2}{\sigma''(t)} \binom{2}{k} \sigma\left(\frac{k}{K}x + t\right). \tag{A.6.3}$$

Since σ is three times continuously differentiable on (a,b) and $(k-1)x/K + t \in (a,b)$ if $x \in [0,1]$, it can be expanded in the Taylor series with Lagrange remainder around t to have

$$\begin{aligned} f_\sigma(x|\boldsymbol{\theta}_2) &= \frac{K^2}{\sigma''(t)} \sum_{k=0}^{2} (-1)^k \binom{2}{k} \\ &\quad \times \left\{ \sigma(t) + \sigma'(t)\frac{kx}{K} + \frac{\sigma''(t)}{2}\frac{(kx)^2}{K^2} + \frac{\sigma''(\xi_k)}{6}\frac{(kx)^3}{K^3} \right\} \\ &= \frac{K^2}{\sigma''(t)} \left\{ \sigma''(t)\frac{x^2}{K^2} + \sum_{k=1}^{2} (-1)^k \binom{2}{k} \frac{\sigma'''(\xi_k)}{6}\frac{(kx)^3}{K^3} \right\} \\ &= x^2 + \frac{x^3}{6K\sigma''(t)} \sum_{k=1}^{2} (-1)^k k^3 \binom{2}{k} \sigma'''(\xi_k), \end{aligned}$$

where $\xi_k \in [t - k|x|/K, t + k|x|/K] \subset (a,b)$. Since the third order derivative is bounded on (a,b), we get the desired assertion by retaking $K \leftarrow \sqrt{2/\sigma''(t)}K$.

Proof of (b). The proof can be done straightforwardly by the polarization type identity:

$$x_1 x_2 = 2A^2 \left\{ \left(\frac{x_1 + x_2}{2A}\right)^2 - \left(\frac{x_1}{2A}\right)^2 - \left(\frac{x_1}{2A}\right)^2 \right\}.$$

We construct the network as

$$f_\sigma(\mathbf{x}|\boldsymbol{\theta}_{\times,A}) := 2A^2\left\{f_\sigma\left(\frac{x_1+x_2}{2A}\middle|\boldsymbol{\theta}_2\right) - f_\sigma\left(\frac{x_1}{2A}\middle|\boldsymbol{\theta}_2\right) - f_\sigma\left(\frac{x_2}{2A}\middle|\boldsymbol{\theta}_2\right)\right\}, \tag{A.6.4}$$

where $\boldsymbol{\theta}_2$ is defined in (A.6.3). Since $(x_1+x_2)/2A, x_1/2A, x_2/2A \in [-1,1]$ for $\mathbf{x} \in [-A,A]^2$, the triangle inequality implies that $|f_\sigma(\mathbf{x}|\boldsymbol{\theta}_{\times,A}) - x_1x_2| \le 6A^2C_1/K$.

Proof of (c). Let $q := \lceil \log_2 \alpha \rceil$. We construct $\boldsymbol{\theta}_{\mathbf{m}}$ as follows. Fix $\mathbf{x} \equiv (x_1,\ldots,x_d) \in [0,1]^d$. We first consider the affine map that transforms $(x_1,\ldots,x_d)$ to $\mathbf{z} \in [0,1]^{2^q}$ which is given by

$$\mathbf{z} := (\underbrace{x_1,\ldots,x_1}_{m_1 \text{ times}}, \underbrace{x_2,\ldots,x_2}_{m_2 \text{ times}}, \ldots, \underbrace{x_d,\ldots,x_d}_{m_d \text{ times}}, \underbrace{1,\ldots,1}_{2^q-|\mathbf{m}| \text{ times}}).$$

The first hidden layer of $\boldsymbol{\theta}_{\mathbf{m}}$ pairs neighboring entries in $\mathbf{z}$ and applies the network $\boldsymbol{\theta}_{\times,A_1}$ defined in (b) with $A_1 = 1$ to each pair. That is, the first hidden layer of $\boldsymbol{\theta}_{\mathbf{m}}$ produces

$$\left\{g_{1,j} := f_\sigma((z_{2j-1},z_{2j})|\boldsymbol{\theta}_{\times,1}) : j = 1,\ldots,2^{q-1}\right\}.$$

Note that $\sup_{1\le j\le 2^{q-1}} |g_{1,j} - z_{2j-1}z_{2j}| \le 6C_1/K$ and $\sup_{1\le j\le 2^{q-1}} |g_{1,j}| \le 6C_1/K + 1$, where $6C_1/K+1$ can be bounded by some constant $A_2 > 1$ depending only on C_1 and K_0. Then the second hidden layer of $\boldsymbol{\theta}_{\mathbf{m}}$ pairs neighboring entries of $\left\{g_{1,j} : j = 1,\ldots,2^{q-1}\right\}$ and applies $\boldsymbol{\theta}_{\times,A_2}$ to each pair to have

$$\left\{g_{2,j} := f_\sigma((g_{1,2j-1},g_{1,2j})|\boldsymbol{\theta}_{\times,A_2}) : j = 1,\ldots,2^{q-2}\right\}.$$

Note that $\sup_{1\le j\le 2^{q-2}} |g_{2,j} - g_{1,2j-1}g_{1,2j}| \le 6C_1A_2^2/K$ and $\sup_{1\le j\le 2^{q-2}} |g_{2,j}| \le 6C_1A_2^2/K + 1 \le A_3$ for some $A_3 > 1$ depending only on C_1 and K_0. We repeat this procedure to produce $\left\{g_{k,j} : j = 1,\ldots,2^{q-k}\right\}$ for $k = 3,\ldots,q$ with

$$\sup_{1\le j\le 2^{q-k}} \left|g_{k,j} - g_{k-1,2j-1}g_{k-1,2j}\right| \le \frac{6C_1A_k^2}{K}, \qquad \sup_{1\le j\le 2^{q-k}} \left|g_{k,j}\right| \le A_{k+1},$$

for some $A_{k+1} > 1$, and we set $f_\sigma(\mathbf{x}|\boldsymbol{\theta}_{\mathbf{m}})$ equal to $g_{q,1}$.

By applying the triangle inequality repeatedly, we have

$$\begin{aligned}
\left|g_{q,1}-\mathbf{x}^{\mathbf{m}}\right| &\le \left|g_{q,1}-g_{q-1,1}g_{q-1,2}\right| + \left|g_{q-1,1}-\prod_{j=1}^{2^{q-1}} z_j\right|\left|g_{q-1,2}\right| \\
&\quad + \left|g_{q-1,2}-\prod_{j=2^{q-1}+1}^{2^q} z_j\right|\left|\prod_{j=1}^{2^{q-1}} z_j\right| \\
&\le \frac{6C_1A_q^2}{K} + A_q\left|g_{q-1,1}-\prod_{j=1}^{2^{q-1}} z_j\right| + \left|g_{q-1,2}-\prod_{j=2^{q-1}+1}^{2^q} z_j\right| \\
&\le \frac{6C_1A_q^2}{K} + (A_q+1)\frac{6C_1A_{q-1}^2}{K} \\
&\quad + A_qA_{q-1}\left|g_{q-2,1}-\prod_{j=1}^{2^{q-2}} z_j\right| + A_q\left|g_{q-2,2}-\prod_{j=2^{q-2}+1}^{2\times 2^{q-2}} z_j\right| \\
&\quad + A_{q-1}\left|g_{q-2,3}-\prod_{j=2\times 2^{q-2}+1}^{3\times 2^{q-2}} z_j\right| + \left|g_{q-2,4}-\prod_{j=3\times 2^{q-2}+1}^{4\times 2^{q-2}} z_j\right| \\
&\le \cdots \le \sum_{k=0}^{q-1}\left\{A_{q-k}^2\prod_{h=q-k+1}^{q}(A_h+1)\right\}\frac{6C_1}{K} \le C_1'\frac{1}{K},
\end{aligned}$$

for some $C_1' > 0$ depending only on C_1, K_0 and q. Since we set $\mathbf{x}$ arbitrary, the proof is done.

Proof of (d). By (b), it is easy to verify that there is a network $\boldsymbol{\theta}_1 \in \Theta_{1,1}(1,6)$ with $|\boldsymbol{\theta}_1|_\infty \le \max\{K^2,2\}$ such that $|\sigma(x)-x| \le C_1'/K$ for any $x \in [-1,1]$ and some constant $C_1' > 0$. The Taylor series with Lagrange remainder around 1 of $\sqrt{x}$ is given by

$$\sqrt{x} = \sum_{k=0}^{J}\frac{(x-1)^k}{k!} + \frac{1}{(J+1)!}\frac{\mathrm{d}^{J+1}\sqrt{x}}{\mathrm{d}x^{J+1}}\Big|_{x=\xi}(x-1)^{J+1},$$

where $\xi \in [0,2]$, and thus

$$\sup_{x\in[0,2]} \left| \sqrt{x} - \sum_{k=0}^{J} \frac{(x-1)^k}{k!} \right| \le C_1' \frac{1}{(J+1)!} \le \mathrm{e} \left(\frac{\mathrm{e}}{J+1} \right)^{J+1}.$$

for some $C_1' > 0$, where the last inequality is because $n! \ge (n/\mathrm{e})^n \mathrm{e}$.

Now, we will construct a neural network $\boldsymbol{\theta}_{p,J}$ that approximates the polynomial $\sum_{k=0}^{J} \frac{(x-1)^k}{k!}$ as follows. The first hidden layer computes $(f_\sigma(x-1|\boldsymbol{\theta}_2)/2, f_\sigma(x-1|\boldsymbol{\theta}_1))$ from the input x. Then

$$\left| \big(f_\sigma(x-1|\boldsymbol{\theta}_2)/2, f_\sigma(x-1|\boldsymbol{\theta}_1)\big) - \Big((x-1)^2/2, (x-1)\Big) \right|_\infty \le C_2' \frac{1}{K},$$

for any $x \in [0,1]$ and some constant $C_2' > 0$. The next hidden layer computes $(f_\sigma((u,v)|\boldsymbol{\theta}_{\times,1+C_2'/K})/3, f_\sigma(u+v|\boldsymbol{\theta}_1))$ from the input (u,v) from the first hidden layer. Using the triangle inequality, we have that the second hidden layer approximates the vector $((x-1)^3/3!, (x-1)^2/2 + (x-1))$ by error $\le 2C_3'/K$ for some $C_3' > 0$. Repeating this procedure, we construct the network $\boldsymbol{\theta}_{p,J} \in \Theta_{1,1}(J, 15)$ which approximates $\sum_{k=0}^{J} \frac{(x-1)^k}{k!}$ by error $\le C_4' J/K$ for some $C_4' > 0$. Taking $J = \lceil \log K \rceil$, we observe that $(\mathrm{e}/J+1)^{J+1} \le (\mathrm{e}/\log K)^{\log K + 1} \le \mathrm{e}K/(\log K)^{\log K} \le 1/K$ for all sufficiently large K, which implies the desired result.

Proof of (e). Let $\zeta \in (0,1)$. Since for any $x \in \mathbb{R}$,

$$\sqrt{x^2+\zeta^2} - |x| \le \frac{\zeta^2}{\sqrt{x^2+\zeta^2}+|x|} \le \frac{\zeta^2}{\zeta} = \zeta,$$

the function $\sqrt{x^2+\zeta^2}$ approximates the absolute value function $|x|$ by error ζ. For $\boldsymbol{\theta}_2$ in (a) and $\boldsymbol{\theta}_{1/2}$ in (d), we have that

$$\begin{aligned}
&\left|f_\sigma\left(f_\sigma(x|\boldsymbol{\theta}_2)+\zeta^2\middle|\boldsymbol{\theta}_{1/2}\right)-|x|\right| \\
&\leq\left|f_\sigma\left(f_\sigma(x|\boldsymbol{\theta}_2)+\zeta^2\middle|\boldsymbol{\theta}_{1/2}\right)-\sqrt{x^2+\zeta^2}\right|+\zeta \\
&\leq\left|f_\sigma\left(f_\sigma(x|\boldsymbol{\theta}_2)+\zeta^2\middle|\boldsymbol{\theta}_{1/2}\right)-\sqrt{f_\sigma(x|\boldsymbol{\theta}_2)+\zeta^2}\right| \\
&\quad+\left|\sqrt{f_\sigma(x|\boldsymbol{\theta}_2)+\zeta^2}-\sqrt{x^2+\zeta^2}\right|+\zeta \\
&\leq C_1'\left(\frac{\log K}{K}+\frac{1}{K\zeta}\right)+\zeta
\end{aligned}$$

for some constant $C_1'>0$. We now set $\zeta=1/\sqrt{K}$ and $f_\sigma(x|\boldsymbol{\theta}_{\text{abs}}):=f_\sigma(f_\sigma(x|\boldsymbol{\theta}_2)+K^{-1}|\boldsymbol{\theta}_{1/2})$. Since $(\log K)/K=o(1/\sqrt{K})$, the proof is done. □

Proof of Theorem A.4.1 for locally quadratic activation functions. Recall that

$$P_M(\mathbf{x})=\sum_{\mathbf{z}\in\mathbb{G}_{d,M}}\sum_{\mathbf{m}\in\mathbb{N}_0^d:|\mathbf{m}|\leq\alpha}\beta_{\mathbf{z},\mathbf{m}}\mathbf{x}^{\mathbf{m}}\phi_{\mathbf{z},M}(\mathbf{x}).$$

Then by Lemma B.1 of [80],

$$\sup_{\mathbf{x}\in[0,1]^d}\left|P_M(\mathbf{x})-f(\mathbf{x})\right|\leq RM^{-\alpha}.$$

From the equivalent representation of the ReLU function $(x)_+=(x+|x|)/2$, we can easily check that the neural network

$$f_\sigma(x|\boldsymbol{\theta}_{\text{relu}}):=\left(f_\sigma(x|\boldsymbol{\theta}_{\text{abs}})+f_\sigma(x|\boldsymbol{\theta}_1)\right)/2$$

with $\boldsymbol{\theta}_{\text{relu}}\in\Theta_{1,1}(\lceil\log K\rceil,21)$ approximates the ReLU function by error $\leq C_1'/\sqrt{K}$ for some $C_1'>0$, where $\boldsymbol{\theta}_1\in\Theta_{1,1}(1,6)$ is defined in the proof of (d) of Lemma A.6.3 and $\boldsymbol{\theta}_{\text{abs}}\in\Theta_{1,1}(\lceil\log K\rceil,15)$ is defined in (e) of Lemma A.6.3. For $z\in(0,1)$ and $M\in\mathbb{N}$, we define

$$f_\sigma(x|\boldsymbol{\theta}_{\phi,z,M}):=f_\sigma\left(1/M-f_\sigma((x-z)|\boldsymbol{\theta}_{\text{abs}})\middle|\boldsymbol{\theta}_{\text{relu}}\right).$$

Then it approximates the function $(1/M - |x-z|)_+$ by error $\leq C_2'/\sqrt{K}$ for some $C_2' > 0$. In turn, for $\mathbf{z} \in \mathbb{G}_{d,M}$, by invoking the similar construction used in (c) of Lemma A.6.3 to approximates the product of d components, we can construct the network $\boldsymbol{\theta}_{\phi,\mathbf{z},M} \in \Theta_{1,1}(\lceil \log K \rceil + \lceil \log_2 d \rceil, 21d)$ with $\left|\boldsymbol{\theta}_{\phi,\mathbf{z},M}\right|_\infty \leq C_3' K^2$ for some $C_3' > 0$ such that

$$\sup_{\mathbf{x}\in[0,1]^d} \left| f_\sigma(\mathbf{x}|\boldsymbol{\theta}_{\phi,\mathbf{z},M}) - \prod_{j=1}^d \left(\frac{1}{M} - |x_j - z_j|\right)_+ \right| \leq C_4' \frac{1}{\sqrt{K}},$$

for some $C_4' > 0$. For each $\mathbf{m} \in \mathbb{N}_0^d$ with $|\mathbf{m}| \leq \alpha$, we have the neural network $\boldsymbol{\theta}_\mathbf{m}$ in (c) of Lemma A.6.3 that approximates $\mathbf{x}^\mathbf{m}$. The number of these networks is $\binom{d+\alpha}{\alpha}$, which is denoted by A_α. Also there are $|\mathbb{G}_{d,M}| = (M+1)^d$ networks $\boldsymbol{\theta}_{\phi,\mathbf{z},M}$ for $\mathbf{z} \in \mathbb{G}_{d,M}$. We need approximation of each product $\mathbf{x}^\mathbf{m}\phi_{\mathbf{z},M}$, which requires additional $A_\alpha(M+1)^d$ many networks $\boldsymbol{\theta}_{\times,A} \in \Theta_{2,1}(1,9)$, where $\boldsymbol{\theta}_{\times,A}$ is defined as in (A.6.4) for some $A > 1$ not depending on M and K. Finally we construct the network $\boldsymbol{\theta}_{f,K,M}$ as

$$f_\sigma(\mathbf{x}|\boldsymbol{\theta}_{f,K,M}) := \sum_{\mathbf{m}\in\mathbb{N}_0^d:|\mathbf{m}|\leq\alpha} \sum_{\mathbf{z}\in\mathbb{G}_{d,M}} \beta_{\mathbf{z},\mathbf{m}} f_\sigma\left((f_\sigma(\mathbf{x}|\boldsymbol{\theta}_\mathbf{m}), f_\sigma(\mathbf{x}|\boldsymbol{\theta}_{\phi,\mathbf{z},M}))|\boldsymbol{\theta}_{\times,A}\right) \tag{A.6.5}$$

Then

$$\begin{aligned}\sup_{\mathbf{x}\in[0,1]^d} \left|f_\sigma(\mathbf{x}|\boldsymbol{\theta}_{f,K,M}) - P_M(\mathbf{x})\right| &\leq C_5' A_\alpha (M+1)^d \left(\frac{1}{K} + \frac{1}{\sqrt{K}}\right) \\ &\leq C_6' \frac{(M+1)^d}{\sqrt{K}},\end{aligned}$$

for some positive constants C_5' and C_6'. In addition, we have $L(\boldsymbol{\theta}_{f,K,M}) \leq 1 + (\lceil \log K \rceil + \lceil \log_2(\alpha \vee d) \rceil \leq C_7' \lceil \log K \rceil$ and $N_{\max}(\boldsymbol{\theta}_{f,K,M}) \leq C_8' A_\alpha (M+1)^d$ for some positive constants C_7' and C_8'. For sparsity of the network, we have

$$\begin{aligned}\left|\boldsymbol{\theta}_{f,K,M}\right|_0 &\leq A_\alpha (M+1)^d \left|\boldsymbol{\theta}_{\times,A}\right|_0 + (M+1)^d \left|\boldsymbol{\theta}_{\phi,\mathbf{z},M}\right|_0 + A_\alpha |\boldsymbol{\theta}_\mathbf{m}|_0 \\ &\leq C_9' \lceil \log K \rceil (M+1)^d,\end{aligned}$$

for some $C_9' > 0$. Taking $M+1 = \epsilon^{-1/\alpha}$ and $K = \epsilon^{-2d/\alpha-2}$, we have

$$\boldsymbol{\theta}_{f,K,M} \in \Theta\left(L_0 \log(1/\epsilon), N_0 \epsilon^{-d/\alpha}, S_0 \epsilon^{-d/\alpha} \log(1/\epsilon), B_0 \epsilon^{-4(d/\alpha+1)}\right),$$

so that $\left\| P_M - f_\sigma(\cdot|\boldsymbol{\theta}_{f,K,M}) \right\|_\infty \leq C_{10}' \epsilon$ for some $C_{10}' > 0$. Since $\left\| f - P_M \right\|_\infty \leq RM^{-\alpha} \leq C_{11}' \epsilon$ for some $C_{11}' > 0$, the proof is done. □

A.6.3 Proof of Proposition A.5.1

Proof. Given a neural network $\boldsymbol{\theta} = ((\mathbf{W}_1, \mathbf{b}_1), \ldots, (\mathbf{W}_{L+1}, \mathbf{b}_{L+1})) \in \Theta_{d,1}(L, N, S, B)$, we define $\check{f}_{l,\sigma,\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^{n_{l-1}}$ and $\hat{f}_{l,\sigma,\boldsymbol{\theta}} : \mathbb{R}^{n_l} \to \mathbb{R}$ as

$$\begin{aligned}
\check{f}_{l,\sigma,\boldsymbol{\theta}}(\mathbf{x}) &:= \sigma_{l-1} \circ \mathsf{A}_{l-1} \circ \cdots \circ \sigma_1 \circ \mathsf{A}_1(\mathbf{x}), \\
\hat{f}_{l,\sigma,\boldsymbol{\theta}}(\mathbf{x}) &:= \mathsf{A}_{L+1} \circ \sigma_L \circ \mathsf{A}_L \circ \cdots \sigma_l \circ \mathsf{A}_l \circ \sigma_{l-1}(\mathbf{x}),
\end{aligned}$$

for $l \in 2, \ldots, L$, where $\mathsf{A}_l \mathbf{x} = \mathbf{W}_l \mathbf{x} + \mathbf{b}_l$. Corresponding to the last and first layer, we define $\check{f}_{1,\sigma,\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}$ and $\hat{f}_{L+1,\sigma,\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}$. Note that $f_\sigma(\mathbf{x}|\boldsymbol{\theta}) = \hat{f}_{l+1,\sigma,\boldsymbol{\theta}} \circ \mathsf{A}_l \circ \check{f}_{l,\sigma,\boldsymbol{\theta}}(\mathbf{x})$. For given $\delta > 0$, let $\boldsymbol{\theta} = ((\mathbf{W}_1, \mathbf{b}_1), \ldots, (\mathbf{W}_{L+1}, \mathbf{b}_{L+1})) \in \Theta_{d,1}(L, N, S, B)$ and $\boldsymbol{\theta}^* = ((\mathbf{W}_1^*, \mathbf{b}_1^*), \ldots, (\mathbf{W}_{L+1}^*, \mathbf{b}_{L+1}^*)) \in \Theta_{d,1}(L, N, S, B)$ be two neural network parameter such that $\left|\mathrm{vec}(\mathbf{W}_l - \mathbf{W}_l^*)\right|_\infty \leq \delta$ and $\left|\mathbf{b}_l - \mathbf{b}_l^*\right|_\infty \leq \delta$ for $l = 1, \ldots, L+1$. Let C_σ be the Lipschitz constant of σ. We observe that

$$\begin{aligned}
\left\| \check{f}_{l,\sigma,\boldsymbol{\theta}} \right\|_\infty &\leq C_\sigma \left(NB \left\| \check{f}_{l-1,\sigma,\boldsymbol{\theta}} \right\|_\infty + B \right) \\
&\leq C_\sigma (B \vee 1)(N+1) \left(\left\| \check{f}_{l-1,\sigma,\boldsymbol{\theta}} \right\|_\infty \vee 1 \right) \\
&\leq \left\{ C_\sigma (B \vee 1)(N+1) \right\}^{l-1} \left(\|\mathbf{x}\|_\infty \vee 1 \right) \\
&\leq \left\{ C_\sigma (B \vee 1)(N+1) \right\}^{l-1},
\end{aligned}$$

and similarly, for any $\mathbf{z}_1 \in \mathbb{R}^N$ and $\mathbf{z}_2 \in \mathbb{R}^N$,

$$\left| \hat{f}_{l+1,\sigma,\boldsymbol{\theta}}(\mathbf{z}_1) - \hat{f}_{l+1,\sigma,\boldsymbol{\theta}}(\mathbf{z}_2) \right| \leq \left\{ C_\sigma (B \vee 1)(N+1) \right\}^{L-l} \|\mathbf{z}_1 - \mathbf{z}_2\|_\infty.$$

Letting $\mathsf{A}_l^*\mathbf{x} = \mathbf{W}_l^*\mathbf{x} + \mathbf{b}_l^*$, we have

$$\begin{aligned}
&\left\|f_\sigma(\cdot|\boldsymbol{\theta}) - f_\sigma(\cdot|\boldsymbol{\theta}^*)\right\|_\infty \\
&\leq \left\|\sum_{l=1}^{L}\left[\hat{f}_{l+1,\sigma,\boldsymbol{\theta}^*} \circ \mathsf{A}_l \circ \check{f}_{l,\sigma,\boldsymbol{\theta}}(\cdot) - \hat{f}_{l+1,\sigma,\boldsymbol{\theta}^*} \circ \mathsf{A}_l^* \circ \check{f}_{l,\sigma,\boldsymbol{\theta}}(\cdot)\right]\right\|_\infty \\
&\leq \sum_{l=1}^{L}(C_\sigma BN)^{L-l}\left\|(\mathsf{A}_l - \mathsf{A}_l^*) \circ \check{f}_{l,\sigma,\boldsymbol{\theta}}(\cdot)\right\|_\infty \\
&\leq \sum_{l=1}^{L}(C_\sigma BN)^{L-l}\delta\left[N\left\{C_\sigma(B\vee 1)(N+1)\right\}^{l-1} + 1\right] \\
&\leq \delta L\left\{C_\sigma(B\vee 1)(N+1)\right\}^{L}.
\end{aligned}$$

Thus, for a fixed sparsity pattern (i.e., the location of nonzero elements in $\boldsymbol{\theta}$), the covering number is bounded by $\left[\delta/L\left\{C_\sigma(B\vee 1)(N+1)\right\}^{L}\right]^{-S}$. Since the number of the sparsity patterns is bounded by $\binom{(N+1)^L}{S} \leq (N+1)^{LS}$, the log of covering number is bounded above by

$$\begin{aligned}
&\log\left((N+1)^{LS}\left[\frac{L\left\{C_\sigma(B\vee 1)(N+1)\right\}^{L}}{\delta}\right]^{S}\right) \\
&\leq 2L(S+1)\log\left(\frac{C_\sigma L(B\vee 1)(N+1)}{\delta}\right),
\end{aligned}$$

which completes the proof. □

A.6.4 Proof of Theorem A.5.2

The proof Theorem A.5.2 is based on the following oracle inequality.

Lemma A.6.4 (Lemma 4 of [80]). *Assume that $Y|\mathbf{X} = \mathbf{x} \sim \mathsf{N}(f_0(\mathbf{x}), 1)$ for some f_0 with $\|f_0\|_\infty \leq R$. Let $\mathcal{F}^\dagger$ be a given function class from $[0,1]^d$ to $[-2R, 2R]$,*

and let $\hat{f}$ be any estimator in $\mathcal{F}^\dagger$. Then for any $\delta \in (0,1]$, we have

$$\begin{aligned}
&\mathsf{E}\left[\mathsf{E}_{\mathbf{X}\sim\mathsf{P}_\mathbf{X}}\left(\hat{f}(\mathbf{X}) - f_0(\mathbf{X})\right)^2\right] \\
&\le 4\Bigg[\Delta_n + \inf_{f\in\mathcal{F}^\dagger} \mathsf{E}_{\mathbf{X}\sim\mathsf{P}_\mathbf{X}}\left(f(\mathbf{X}) - f_0(\mathbf{X})\right)^2 \\
&\qquad + (4R)^2\frac{18\log\mathcal{N}(\delta,\mathcal{F}^\dagger,\|\cdot\|_\infty)+72}{n} + 32\delta(4R)\Bigg],
\end{aligned}$$

with

$$\Delta_n := \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^n\left(Y_i - \hat{f}\left(\mathbf{X}_i\right)\right)^2 - \inf_{f\in\mathcal{F}^\dagger}\frac{1}{n}\sum_{i=1}^n\left(Y_i - f\left(\mathbf{X}_i\right)\right)^2\right],$$

where the expectations are taken over the training data.

Proof of Theorem A.5.2. We apply Lemma A.6.4 to $\mathcal{F}^\dagger = \mathcal{F}_{\sigma,n}$ and

$$\hat{f} = \hat{f}_n \in \operatorname*{argmin}_{f\in\mathcal{F}_{\sigma,n}}\sum_{i=1}^n\left(Y_i - f(\mathbf{X}_i)\right)^2.$$

By definition of $\hat{f}_n$, we have $\Delta_n = 0$. Set $\delta = 1/n$. By Proposition A.5.1,

$$\log\mathcal{N}\left(\frac{1}{n},\mathcal{F}_{\sigma,n},\|\cdot\|_\infty\right) \le C_1' n^{\frac{d}{2\alpha+d}}\log^3 n,$$

for some $C_1' > 0$. If a function f_n is approximates f_0 by error ϵ which is sufficiently small, then $\|f_n\|_\infty \le 2R$ since $\|f_0\|_\infty \le R$. Now, Theorem A.4.1 implies that there is $f_n \in \mathcal{F}_{\sigma,n}$ such that

$$\begin{aligned}
\mathsf{E}_{f_0,\mathsf{P}_\mathbf{X}}\left(f_n(\mathbf{X}) - f_0(\mathbf{X})\right)^2 &\le C_2' \sup_{\mathbf{x}\in[0,1]^d}\left|f_n(\mathbf{x}) - f_0(\mathbf{x})\right|^2 \\
&\le C_3'\left(\left(n^{\frac{d}{2\alpha+d}}\right)^{-d/\alpha}\right)^2 = C_3' n^{-\frac{2\alpha}{2\alpha+d}},
\end{aligned}$$

which completes the proof. □

A.6.5 Proof of Theorem A.5.3

For a given real-valued function f, let $\mathcal{R}_{\mathsf{hinge}}(f) := \mathsf{E}\ell_{\mathsf{hinge}}(Yf(\mathbf{X}))$, which we call the hinge risk. The proof of Theorem A.5.3 is based on Theorem 1.7.5.

Proof of Theorem A.5.3. It is well known that

$$f^* = 2\mathbb{1}\left(\eta(\cdot) \geq 1/2\right) - 1 = \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{R}_{\mathsf{hinge},\eta}(f),$$

i.e., the hinge risk minimizer is equal to the Bayes classifier [60]. The first step is to find a function $f_n \in \mathcal{F}_{\sigma,n}$ which approximates the Bayes classifier f^* well. Let $(\xi_n)_{n\in\mathbb{N}}$ be a given sequence of positive integers. Since $\eta \in \mathcal{H}^{\alpha,R}([0,1]^d)$, by Theorem 1.7.5, for each ξ_n there exists $\boldsymbol{\theta}_n$ such that $\|f_\sigma(\cdot|\boldsymbol{\theta}_n) - \eta(\cdot)\|_\infty \leq \xi_n$ with at most $O(\log(1/\xi_n))$ layers, $O(\xi_n^{-d/\alpha})$ nodes at each layer and $O(\xi_n^{-d/\alpha}\log(1/\xi_n))$ nonzero parameters. Then as we did in the proof of Theorem 1.3.1, we can construct f_n by adding one ReLU layer to $f_\sigma(\cdot|\boldsymbol{\theta}_n)$, which satisfies

$$\mathcal{R}_{\mathsf{hinge},\eta}(f_n) - \mathcal{R}_{\mathsf{hinge},\eta}(f^*) \leq C_1'\xi_n^{q+1},$$

for some $C_1' > 0$,

We take $\delta_n = C_1'\xi_n^{q+1}$. Then there are positive constants L_0, N_0, S_0 and B_0 such that $f_n \in \mathcal{F}_{\sigma,n}$ where

$$\mathcal{F}_{\sigma,n} := \Big\{ f_\sigma(\cdot|\boldsymbol{\theta}) : \big\|f_\sigma(\cdot|\boldsymbol{\theta})\big\|_\infty \leq 1, \\ \boldsymbol{\theta} \in \Theta_{d,1}\left(L_0\log(\delta_n^{-1}), N_0\delta_n^{-\frac{d}{\alpha(q+1)}}, S_0\delta_n^{-\frac{d}{\alpha(q+1)}}\log(\delta_n^{-1}), B_0\delta_n^{-\kappa'}\right)\Big\},$$

for some $\kappa' > 0$. Proposition A.5.1 implies that the log covering number of $\mathcal{F}_{\sigma,n}$ is bounded above by

$$\log\mathcal{N}\left(\delta_n, \mathcal{F}_{\sigma,n}, \|\cdot\|_\infty\right) \leq \delta_n^{-d/\alpha(q+1)}\log^3(\delta_n^{-1}).$$

Note that to satisfy the entropy condition (A5), δ_n should satisfy

$$(\delta_n)^{\frac{d}{\alpha(q+1)}+\frac{q+2}{q+1}} \geq C_2' n^{-1}\log^3(\delta_n^{-1}) \tag{A.6.6}$$

for some $C_2' > 0$. If we let $\delta_n = (\log^3 n/n)^{\alpha(q+1)/(\alpha(q+2)+d)}$, the condition (A.6.6) holds and so the proof is done. □

Appendix B

Poisson mixture of finite feature models

B.1 Overview

A latent feature model generates a random binary matrix with a finite number of rows, say p and an infinite number of columns. Each row of the $p \times \infty$ dimensional binary matrix Ξ represents an object and each column represents an unobserved property called a feature. The $\xi_{jk} = 1$ means that the j-th object possesses the k-th feature and $\xi_{jk} = 0$ means that it does not, where ξ_{jk} denotes the (j,k)-th entry of Ξ.

The Indian buffet process (IBP) [38] is a distribution that is popularly used for modeling the latent feature model. The IBP and its two- and three-parameter generalizations [92, 89] has been widely applied to machine learning problems [e.g., 66, 70, 68, 14].

In this chapter we focus on the two-parameter IBP. Construction of the two-parameter IBP starts with the finite feature model which is the distribution on the $p \times K$ dimensional binary matrix Ξ such that

$$\begin{aligned} \theta_k &\overset{\text{iid}}{\sim} \mathsf{Beta}\left(\frac{\alpha}{K}, \kappa + 1\right),\ k \in [K] \\ \xi_{jk} | \theta_k &\overset{\text{ind}}{\sim} \mathsf{Bernoulli}(\theta_k),\ j \in [p], k \in [K] \end{aligned} \tag{B.1.1}$$

for $\alpha > 0$ and $\kappa \geq 0$. The IBP with parameters α and κ, denoted by $\mathsf{IBP}(\alpha, \kappa)$, is the limit in distribution of the finite feature model (B.1.1) as $K \to \infty$.

In this section, we consider another distribution on a binary matrix with an infinite number of columns, called PFM, which is an abbreviation for Poisson mixture of finite feature models. The PFM also starts with the finite feature model, but instead of taking a limit as $K \to \infty$, it imposes the Poisson distribution on the number of features. We say that a $p \times \infty$ dimensional binary matrix Ξ follows $\mathsf{PFM}(\gamma, \alpha, \kappa)$ for $\gamma > 0$, $\alpha > 0$ and $\kappa \geq 0$, if

$$\begin{aligned} K &\sim \mathsf{Poisson}(\gamma), \\ \theta_k &\overset{\text{iid}}{\sim} \mathsf{Beta}\left(\alpha, \kappa + 1\right), \ k \in [K] \\ \xi_{jk} | \theta_k &\overset{\text{ind}}{\sim} \mathsf{Bernoulli}(\theta_k) \ j \in [p], k \in [K]. \end{aligned} \tag{B.1.2}$$

In this chapter, we provide two different probabilistic representations of $\mathsf{PFM}(\gamma, \alpha, \kappa)$, which are useful to construct an posterior computation algorithm. Next, as an application, we use the PFM as the prior distribution on the factor loading matrix for Bayesian estimation of a sparse factor model. We derive the posterior consistency of the factor dimensionality and the near-optimal posterior contraction rate for the covariance matrix.

B.1.1 Equivalence classes

The latent feature model considers the ordering of the features irrelevant. Hence we say that two $p \times \infty$ dimensional binary matrices are equivalent if they are identical up to a permutation of columns. It is convenient to choose a representative of every equivalence class by the *left-ordering procedure*. The left-ordering procedure maps each $p \times \infty$ dimensional binary matrix to its left-ordered version whose columns are ordered by the score s_k defined by

$$s_k := \sum_{j=1}^{p} \xi_{jk} 2^{p-j}$$

i.e., ordered so that $s_1 \geq s_2 \geq \cdots$. We call the equivalence class defined by the left-ordering procedure *lof-equivalence* class and we denote the lof-equivalence class of a binary matrix Ξ by $[\Xi]$.

B.1.2 Notation

We denote by $\mathbb{1}(\cdot)$ the indicator function. Let $\mathbb{R}$ be the set of real numbers and $\mathbb{R}_+$ be the set of positive numbers. Let $\mathbb{N}$ be the set of natural numbers. For $m \in \mathbb{N}$, we let $[m] := \{1, \ldots, m\}$.

Let $\mathsf{B}(a, b)$ denotes the beta function with parameters a and b. Let $\bar{\mathsf{B}}^{a_2,b_2}_{a_1,b_1}$ be the ratio of two beta functions defined as

$$\bar{\mathsf{B}}^{a_2,b_2}_{a_1,b_1} := \frac{\mathsf{B}(a_1 + a_2, b_1 + b_2)}{\mathsf{B}(a_1, b_1)}.$$

B.2 Equivalent representations

In this section, we provide two equivalent probabilistic representations of the PFM.

B.2.1 Urn schemes

Urn schemes generate each row of $\boldsymbol{\Xi}$ conditionally on the previous ones. When we describe urn schemes, we frequently use the following quantity

$$m_{j,k} := \sum_{h=1}^{j} \xi_{hk} \tag{B.2.1}$$

for $j \in [p]$. Let $\Delta := \{0, 1\}^p$ which is a set of p-dimensional binary vectors and $\Delta_1 := \Delta \setminus \{\mathbf{0}\}$ where $\mathbf{0}$ is the vector or zero. For each $\mathbf{u} \in \Delta_1$, we define

$$K_{\mathbf{u}} := \sum_{k=1}^{\infty} \mathbb{1}(\boldsymbol{\xi}_{\cdot k} = \mathbf{u}) \tag{B.2.2}$$

where $\boldsymbol{\xi}_{\cdot k}$ denotes the k-th column of $\boldsymbol{\Xi}$. In words, $K_{\mathbf{u}}$ is the number of columns equal to the binary vector $\mathbf{u}$. We let K^+ be the number of nonzero

columns of Ξ, that is

$$K^+ := \sum_{k=1}^{\infty} \mathbb{1}(\|\boldsymbol{\xi}_{\cdot k}\|_0 > 0) = \sum_{\mathbf{u}\in\Delta_1} K_{\mathbf{u}}.$$

We first provide the probability of a particular lof-equivalence class.

Proposition B.2.1. *If a $p \times \infty$-dimensional random binary matrix $\Xi \equiv (\xi_{jk})_{j\in[p],k\in\mathbb{N}}$ is distributed as $\mathsf{PFM}(\gamma, \alpha, \kappa)$, then*

$$\mathsf{P}([\Xi]) = \frac{\gamma^{K^+}}{\prod_{\mathbf{u}\in\Delta_1} K_{\mathbf{u}}!} e^{-\gamma \sum_{j=1}^{p} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{1,j-1}} \left[\prod_{k=1}^{K^+} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p,k},p-m_{p,k}}\right], \tag{B.2.3}$$

where $K^+ := \sum_{\mathbf{u}\in\Delta_1} K_{\mathbf{u}}$.

From Proposition B.2.1, we can derive an urn scheme of the PFM.

Proposition B.2.2. *The probability distribution defined in Equation (B.2.3) can be derived from the following procedure:*

1. *The first customer tries* $\mathsf{Poisson}(\gamma \mathsf{B}(\alpha+1, \kappa+1)/\mathsf{B}(\alpha, \kappa+1))$ *dishes.*
2. *For every $j \geq 2$, the $(j+1)$-th customer*
   - *tries each previously tasted dish independently according to*

     $$\mathsf{Bernoulli}\left(\frac{m_{j,k}+\alpha}{j+1+\kappa+\alpha}\right) \tag{B.2.4}$$

     where $m_{j,k}$ is the number of people (before $(j+1)$-th customer) who have tried dish k;
   - *and tries*

     $$\mathsf{Poisson}\left(\gamma \frac{\mathsf{B}(\alpha+1, \kappa+j+1)}{\mathsf{B}(\alpha, \kappa+1)}\right) \tag{B.2.5}$$

     new dishes

Proof. See Section B.4.1. □

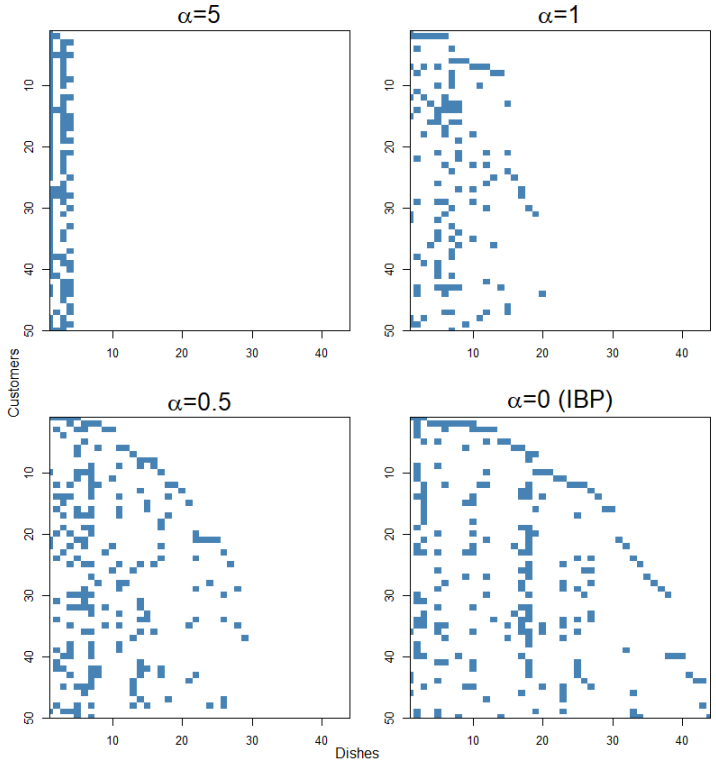


FIGURE B.1: Draws from $\mathsf{PFM}(\omega\kappa/\alpha, \alpha, \kappa)$ and $\mathsf{IBP}(\omega, \kappa)$ with $\gamma = 5$, $\kappa = 4$ but with $\alpha = 5$, $\alpha = 1$ and $\alpha = 0.5$.

The urn scheme of the two-parameter Indian buffet process $\mathsf{IBP}(\omega, \kappa)$ is as follows: The first customer tries $\mathsf{Poisson}(\omega)$ dishes. Then the $(j+1)$-th customer tries each previously tasted dish independently according to $\mathsf{Bernoulli}(m_{j,k}/(j+\kappa+1))$ and tries $\mathsf{Poisson}(\omega\kappa/(j+\kappa+1))$ new dishes.

By comparing this urn scheme with the one of $\mathsf{PFM}(\gamma, \alpha, \kappa)$, it is easy to see that $\mathsf{IBP}(\omega, \kappa)$ is the limit in distribution of $\mathsf{PFM}(\gamma, \alpha, \kappa)$ as $\alpha \to 0$ and $\gamma\alpha/\kappa \to \omega$. Figure B.1 shows four binary matrices generated by $\mathsf{PFM}(\omega\kappa/\alpha, \alpha, \kappa)$ and $\mathsf{IBP}(\omega, \kappa)$ with $\omega = 5$, $\kappa = 4$ but with $\alpha = 5$, $\alpha = 1$ and $\alpha = 0.5$. We can see that the IBP generates more features than the PFM.

B.2.2 Hierarchical representation

We briefly review completely random measures. Let $(\Omega, \mathcal{A})$ a Polish space with its Borel σ-field and let $(\mathfrak{M}, \mathcal{M})$ be a set of all measures on $(\Omega, \mathcal{A})$ with

its Borel σ-field.

A completely random measure (CRM) μ on $(\Omega, \mathcal{A})$ is a random measure such that $\mu(A_1), \ldots, \mu(A_k)$ for all disjoint measurable sets $A_1, \ldots, A_k \in \mathcal{A}$ are mutually independent. Every CRM can be decomposed into three independent parts:

$$\mu = \mu_0 + \sum_{k=1}^{K} q_k \delta_{\omega_k} + \sum_{(q,\omega)\in\Phi} q\delta_\omega$$

where μ_0 is a non-random measure, $(\omega_k)_{k\in[K]}$ are fixed atoms in Ω, $(q_k)_{k\in[K]}$ are independent random variables on $\mathbb{R}_+$ and Φ is a Poisson process on $\mathbb{R}_+ \times \Omega$.

Here we only consider purely-atomic CRMs such that $\mu_0 = 0$. We write

$$\mu \sim \mathsf{CRM}\left(\Lambda, (\omega_k, P_k)_{k\in[K]}\right)$$

if μ is the purely-atomic CRM represented by $\mu = \sum_{k=1}^{K} q_k \delta_{\omega_k} + \sum_{(q,\omega)\in\Phi} q\delta_\omega$ with $q_k \overset{\text{ind}}{\sim} P_k$ for $k \in [K]$ and $\mathsf{E}\Phi = \Lambda$ for some probability measures $(P_k)_{k\in[K]}$ on $\mathbb{R}_+$ and Λ on $\mathbb{R}_+ \times \Omega$. In particular, we write $\mu \sim \mathsf{CRM}(\Lambda)$ if $\mu = \sum_{(q,\omega)\in\Phi} q\delta_\omega$ with $\mathsf{E}\Phi = \Lambda$.

It is well known that the two-parameter Indian buffet process $\mathsf{IBP}(\alpha, \kappa + 1)$ with $\alpha > 0$ and $\kappa \geq 0$ have the following hierarchical representations:

$$\begin{aligned} \boldsymbol{\xi}_{j\cdot} | \mu &\overset{\text{iid}}{\sim} \mathsf{BeP}(\mu),\ j \in [p] \\ \mu &\sim \mathsf{BP}(\kappa + 1, \alpha\Lambda_0) \end{aligned}$$

for some smooth probability measure Λ_0, i.e., $\Lambda_0(\Omega) = 1$. Here, $\mathsf{BeP}(\mu)$ denotes the Bernoulli process with mean μ, which is equivalent to $\mathsf{CRM}(\Lambda_{\mathsf{BeP}(\mu)})$ on $(\Omega, \mathcal{A})$ with

$$\Lambda_{\mathsf{BeP}(\mu)}(\mathrm{d}q, \mathrm{d}\omega) = \delta_1(\mathrm{d}q)\mu(\mathrm{d}\omega),$$

where δ_1 denotes a point mass at 1, and $\mathsf{BP}(\kappa+1, \alpha\Lambda_0)$ denotes the Beta process which is equivalent to $\mathsf{CRM}(\Lambda_{\mathsf{BP}(\theta,\gamma\Lambda_0)})$ on $(\Omega, \mathcal{A})$ with

$$\Lambda_{\mathsf{BP}(\kappa+1,\alpha\Lambda_0)}(\mathrm{d}q, \mathrm{d}\omega) = \alpha(\kappa+1)q^{-1}(1-q)^{\kappa}\mathrm{d}q\Lambda_0(\mathrm{d}\omega).$$

The next theorem provides random measure representation for $\mathsf{PFM}(\gamma, \alpha, \kappa)$.

Proposition B.2.3. *Suppose that*

$$\begin{aligned} \boldsymbol{\xi}_{j\cdot}|\mu &\overset{\text{iid}}{\sim} \mathsf{BeP}(\mu),\ j \in [p] \\ \mu &\sim \mathsf{CRM}(\Lambda_{\mathsf{PFM}(\gamma,\alpha,\kappa)}) \end{aligned} \tag{B.2.6}$$

with

$$\Lambda_{\mathsf{PFM}(\gamma,\alpha,\kappa)}(\mathrm{d}q, \mathrm{d}\omega) = \frac{\gamma}{\mathsf{B}(\alpha, \kappa+1)} q^{\alpha-1}(1-q)^{\kappa}\mathrm{d}q\Lambda_0(\mathrm{d}\omega) \tag{B.2.7}$$

for some smooth probability measure Λ_0. Then $\boldsymbol{\Xi} \equiv (\boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p\cdot})^{\top} \sim \mathsf{PFM}(\gamma, \alpha, \kappa)$.

Proof. See Section B.4.1. □

The function $q \mapsto q^{\alpha-1}(1-q)^{\kappa}$ is integrable, which means that there would be a finite number of features.

B.3 Application to sparse Bayesian factor models

In this section, we consider an application of the PFM distribution to Bayesian estimation of the factor model.

B.3.1 Model and prior

We consider the factor model given by

$$\mathbf{Y}|\mathbf{Z} = \mathbf{z} \sim \mathsf{N}_p(\mathbf{B}\mathbf{z}, \sigma^2\mathbf{I}), \quad \mathbf{Z} \sim \mathsf{N}_K(\mathbf{0}, \mathbf{I}) \tag{B.3.1}$$

where $\mathbf{B}$ is a $p \times K$ factor loading matrix, $\mathbf{Z}$ is a K-dimensional factor with $K < p$, and $\sigma^2 > 0$ is a noise variance.

We consider the following prior on the loading matrix $\mathbf{B}$. Let β_{jk} be the (j,k)-th entry of the $p \times \infty$-dimensional loading matrix $\mathbf{B}$. We impose the prior distribution based on the PFM distribution such that

$$\begin{aligned}
\beta_{jk}|\xi_{jk} &\overset{\text{ind}}{\sim} (1-\xi_{jk})\delta_0 + \xi_{jk}\mathsf{Laplace}(1),\ j \in [p_n],\ k \in [K] \\
\xi_{jk}|\theta_k &\overset{\text{ind}}{\sim} \mathsf{Bernoulli}\left(\theta_k\right),\ j \in [p_n],\ k \in [K] \\
\theta_k &\overset{\text{iid}}{\sim} \mathsf{Beta}(1, \kappa+1),\ k \in [K] \\
K &\sim \mathsf{Poisson}(\gamma)
\end{aligned}$$

where $\alpha > 0$, $\kappa \geq 0$ and $\gamma > 0$ are hyperparameters. We refer to the above distribution on $\mathbf{B}$ as $\mathsf{SSPFM}_p(\gamma, \kappa)$, which is an abbreviation of *spike and slab Poisson mixture of finite feature models*. We denote by $\Pi(\cdot)$ the prior distribution of $\mathsf{SSPFM}_p(\gamma, \kappa)$.

B.3.2 Assumptions on the true distribution

We assume that we observe the data from the model

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{\text{iid}}{\sim} \mathsf{N}_{p_n}(\mathbf{0}, \mathbf{\Sigma}_{0n})$$

where $\mathbf{\Sigma}_{0n}$ is of the form

$$\mathbf{\Sigma}_{0n} := \mathbf{B}_{0n}\mathbf{B}_{0n}^\top + \sigma_{0n}^2\mathbf{I}$$

and $\mathbf{B}_{0n} \in \mathbb{R}^{p_n \times k_{0n}}$. Here, k_{0n} is equivalent to the true factor dimensionality.

We introduce some regularity conditions on the sequence of the true covariance matrices $\{\mathbf{\Sigma}_{0n}\}_{n\in\mathbb{N}}$. Assume that there exist sequences of positive real numbers $\{c_n\}_{n\in\mathbb{N}}$, $\{s_n\}_{n\in\mathbb{N}}$ satisfying the following conditions:

(A1) $\sum_{j=1}^{p_n} \mathbb{1}\left(\sum_{k=1}^{k_{0n}} |\beta_{0n,jk}| > 0\right) \leq s_n$, where $\beta_{0n,jk}$ denotes the (j,k)-th entry of $\mathbf{B}_{0n}$.

(A2) $\|\mathbf{\Sigma}_{0n}\| = c_n \lesssim s_n$.

(A3) $c_0 \leq \sigma_{0n}^2 \leq c_n$ for an universal constant $c_0 > 0$.

(A4) $c_n^2 s_n k_{0n}^2 \log p_n / n = o(1)$, $1 \leq k_{0n} < p_n/2$ and $p_n > n$.

(A5) $\lambda_{k_{0n}}\left(\mathbf{B}_{0n}\mathbf{B}_{0n}^\top\right) > d_0$ for an universal constant $d_0 > 0$.

The above assumptions are the same as Chapter 3.

Given data $\mathbf{Y}_{1:n} \equiv (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, let $\sigma(\mathbf{Y}_{1:n})$ be the σ-field generated by $\mathbf{Y}_{1:n}$. For a given sample size n and covariance matrix $\mathbf{\Sigma}$, we let $\mathsf{P}_{\mathbf{\Sigma}}$ and $\mathsf{E}_{\mathbf{\Sigma}}$ denote the probability measure and the expectation operator under the law $(\mathsf{N}(\mathbf{0}, \mathbf{\Sigma}))^n$, where we suppress the dependence on n for simplicity.

B.3.3 Preliminary results

In this section, we provide useful properties of the SSPFM prior distribution for asymptotic analysis.

The next result gives a upper bound of the tail probability of the factor dimensionality.

Lemma B.3.1. *If* $\mathbf{B} \sim \mathsf{SSPFM}_p(\gamma, \kappa)$ *with* $\gamma > 0$ *and* $\kappa \geq 0$, *then for any* $k \in \mathbb{N}$,

$$\Pi(\mathsf{K}^+(\mathbf{B}) > k) \leq \gamma^{k+1}. \tag{B.3.2}$$

Proof. Since $K \sim \mathsf{Poisson}(\gamma)$, we have that

$$\begin{aligned}\Pi(\mathsf{K}^+(\mathbf{B}) > k) \leq \Pi(K > k) &= \sum_{h=k+1}^{\infty} \frac{\mathrm{e}^{-\gamma}\gamma^h}{h!} \\ &\leq \sum_{h=k+1}^{\infty} \gamma^h \leq \gamma^{k+1}\end{aligned}$$

which completes the proof. □

Recall the definition of the row support up to k-th column of $\mathbf{B}$ (see Definition 3.2.2):

$$\mathsf{supp}_k(\mathbf{B}) := \left\{ j \in [p] : \sum_{h=1}^{k} |\beta_{jh}| > 0 \right\}.$$

The following lemma shows that SSPFM prior with large κ has an exponential tail bound for $|\mathsf{supp}_k(\mathbf{B})|$.

Lemma B.3.2. *If* $\mathbf{B} \sim \mathsf{SSPFM}_p(\gamma, p^2)$ *for* $\gamma > 0$. *Then for any* $k \in \mathbb{N}$ *satisfying* $k \leq p^{1/3}$ *and* $t > 0$,

$$\mathsf{P}\left(|\mathsf{supp}_k(\mathbf{B})| > t\right) \leq \exp\left(-C_1 t \log p\right)$$

for some universal constant $C_1 > 0$.

Proof. See Section B.4.2. □

We now show that the proposed prior puts sufficiently large mass near the truth. We let $\Pi_n(\cdot)$ be the prior distribution of $\mathsf{SSPFM}_{p_n}(\gamma_n, \kappa_n)$ with data-dependent hyperparameters γ_n and κ_n.

Lemma B.3.3. *Suppose that* $\mathbf{B}_{0n} \in \mathbb{R}^{p_n \times k_{0n}}$ *be a* s_n*-sparse up to* k_{0n} *loading matrix. Assume (A2). Let* $\mathbf{B} \sim \mathsf{SSPFM}_{p_n}(\gamma_n, p_n^2)$ *with* $\gamma_n \lesssim 1$. *Then for any* $n \in \mathbb{N}$ *and* $\eta_n > 0$ *such that* $\eta_n \lesssim 1$,

$$\Pi_n\left(\|\mathbf{B} - \mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n\right) \geq \gamma_n^{k_{0n}} \exp\left(-C_1 s_n k_{0n} \log\left(p_n \vee \eta_n^{-1}\right)\right), \quad \text{(B.3.3)}$$

for some universal constant $C_1 > 0$.

Proof. See Section B.4.2. □

The prior concentration near the truth leads to the following useful bound.

Lemma B.3.4. *Suppose that* $\mathbf{\Sigma}_{0n}$ *satisfies (A1)-(A4). Let* $\mathbf{B} \sim \mathsf{SSPFM}_{p_n}(\gamma_n, p_n^2)$ *for and* $\sigma^2 \sim \mathsf{IG}(a, b)$ *with* $\gamma_n \lesssim 1$ *for* $\delta > 0$, $a > 0$ *and* $b > 0$. *Then there is a Borel measurable set* $\mathfrak{A}_n \in \sigma(\mathbf{Y}_{1:n})$ *with* $\mathsf{P}_{\mathbf{\Sigma}_{0n}}(\mathfrak{A}_n) \leq C_1 / \log n$ *for some universal constant* $C_1 > 0$, *on which*

$$D_n := \int \prod_{i=1}^{n} \frac{f_{\mathbf{\Sigma}}(\mathbf{Y}_i)}{f_{\mathbf{\Sigma}_{0n}}(\mathbf{Y}_i)} \mathrm{d}\Pi_n(\mathbf{\Sigma}) \geq \gamma_n^{k_{0n}} \mathrm{e}^{-C_2 s_n k_{0n} \log p_n}. \quad \text{(B.3.4)}$$

for some universal constant $C_2 > 0$ *depending only on a and b.*

Proof. By using similar arguments used in the proof of Corollary 3.2.4, we have

$$\Pi_n\left(\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0n}\|_{\mathrm{F}} \leq \sqrt{\frac{s_n k_{0n}}{n}}\right) \geq \gamma_n^{k_{0n}} \mathrm{e}^{-C_1 s_n k_{0n} \log p_n}.$$

Then Lemma 3.7.4 in Chapter 3 leads to the desired result. □

B.3.4 Asymptotic properties

In this section, we show that the PFM prior leads to the posterior consistency of the factor dimensionality and the near-optimal posterior contraction of the covariance matrix, as the SSIBP prior considered in Chapter 3 did.

We impose the same assumptions as Chapter 3 on the true data generating distributions. We let

$$\mathcal{C}_{0n} := \left\{\boldsymbol{\Sigma}_{0n} \equiv \mathbf{B}_{0n}\mathbf{B}_{0n}^{\top} + \sigma_{0n}^2\mathbf{I} : \text{ (A1)- (A4) are satisfied}\right\}$$
$$\mathcal{C}_{0n}^{\star} := \left\{\boldsymbol{\Sigma}_{0n} \equiv \mathbf{B}_{0n}\mathbf{B}_{0n}^{\top} + \sigma_{0n}^2\mathbf{I} : \text{ (A1)- (A5) are satisfied}\right\}$$

Theorem B.3.5. *A priori, let* $\mathbf{B} \sim \mathsf{SSPFM}_{p_n}(p_n^{-As_n^2}, p_n^{1+\delta})$ *for sufficiently large* $A > 0$ *and* $\sigma^2 \sim \mathsf{IG}(a, b)$ *for* $\delta > 0$, $a > 0$ *and* $b > 0$. *Then*

$$\sup_{\boldsymbol{\Sigma}_{0n} \in \mathcal{C}_{0n}^{\star}} \mathsf{E}_{\boldsymbol{\Sigma}_{0n}}\left[\Pi_n\left(\mathsf{K}^{+}(\mathbf{B}) \neq k_{0n} \middle| \mathbf{Y}_{1:n}\right)\right] = o(1),$$
$$\sup_{\boldsymbol{\Sigma}_{0n} \in \mathcal{C}_{0n}} \mathsf{E}_{\boldsymbol{\Sigma}_{0n}}\left[\Pi_n\left(\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0n}\| > M c_n \sqrt{\frac{s_n^2 k_{0n} \log p_n}{n}} \middle| \mathbf{Y}_{1:n}\right)\right] = o(1),$$

for sufficiently large $M > 0$, *where* $\Pi_n(\cdot|\mathbf{Y}_{1:n})$ *denotes the posterior distribution induced by the prior* Π_n *and the data* $\mathbf{Y}_{1:n}$.

Proof. See Section B.4.3. □

B.4 Proofs

B.4.1 Proofs of results in Section B.2

Proof of Proposition B.2.1. Recall that $m_{p,k} := \sum_{j=1}^{p} \xi_{jk}$. Given K such that $K \geq K^+$, we have that

$$\begin{aligned}
\mathsf{P}(\Xi|K) &= \prod_{k=1}^{K} \frac{\mathsf{B}(m_{p,k}+\alpha, p-m_{p,k}+\kappa+1)}{\mathsf{B}(\alpha,\kappa+1)} \\
&= \left(\frac{\mathsf{B}(\alpha,p+\kappa+1)}{\mathsf{B}(\alpha,\kappa+1)}\right)^{K-K^+} \prod_{k=1}^{K^+} \frac{\mathsf{B}(m_{p,k}+\alpha, p-m_{p,k}+\kappa+1)}{\mathsf{B}(\alpha,\kappa+1)} \\
&= \left(\bar{\mathsf{B}}^{0,p}_{\alpha,\kappa+1}\right)^{K-K^+} \prod_{k=1}^{K^+} \bar{\mathsf{B}}^{m_{p,k},p-m_{p,k}}_{\alpha,\kappa+1}
\end{aligned}$$

where the second equality follows from reordering the columns such that $m_{p,k} < 0$ if $k \leq K^+$ and $m_{p,k} = 0$ otherwise. Recall that $\bar{\mathsf{B}}^{a_2,b_2}_{a_1,b_1} = \mathsf{B}(a_1+a_2, b_1+b_2)/\mathsf{B}(a_1,b_1)$. If $K < K^+$, $\mathsf{P}(\Xi|K) = 0$. The probability of a lof equivalence class of Ξ is

$$\mathsf{P}([\Xi]|K) = \frac{K!}{\prod_{\mathbf{u}\in\Delta} K_{\mathbf{u}}!} \left(\bar{\mathsf{B}}^{0,p}_{\alpha,\kappa+1}\right)^{K-K^+} \prod_{k=1}^{K^+} \bar{\mathsf{B}}^{m_{p,k},p-m_{p,k}}_{\alpha,\kappa+1}$$

Let $p_K(k) := \mathrm{e}^{-\gamma}\gamma^k/k!$ which is the probability mass function of $\mathsf{Poisson}(\gamma)$. Marginalizing over K, we have that

$$\mathsf{P}([\Xi]) = \frac{1}{\prod_{\mathbf{u}\in\Delta_1} K_{\mathbf{u}}!} \left[\prod_{k=1}^{K^+} \bar{\mathsf{B}}^{m_{p,k},p-m_{p,k}}_{\alpha,\kappa+1}\right] \sum_{K=K^+}^{\infty} \frac{K!}{K_{\mathbf{0}}!} \left(\bar{\mathsf{B}}^{0,p}_{\alpha,\kappa+1}\right)^{K-K^+} p_K(K)$$

where the summation term can be written as

$$\begin{aligned}
\sum_{K=K_+}^{\infty} \frac{K!}{K_{\mathbf{0}}!} \left(\bar{\mathsf{B}}^{0,p}_{\alpha,\kappa+1}\right)^{K-K^+} p_K(k) &= \mathrm{e}^{-\gamma}\gamma^{K^+} \sum_{K=K_+}^{\infty} \frac{1}{(K-K^+)!} \left(\gamma\bar{\mathsf{B}}^{0,p}_{\alpha,\kappa+1}\right)^{K-K^+} \\
&= \gamma^{K^+} \mathrm{e}^{-\gamma\left(1-\bar{\mathsf{B}}^{0,p}_{\alpha,\kappa+1}\right)}.
\end{aligned} \tag{B.4.1}$$

From the identity $\mathsf{B}(x,y) - \mathsf{B}(x,y+1) = \mathsf{B}(x+1,y)$, it follows that

$$\begin{aligned} 1 - \frac{\mathsf{B}(\alpha, p+\kappa+1)}{\mathsf{B}(\alpha,\kappa+1)} &= \frac{1}{\mathsf{B}(\alpha,\kappa+1)} \left\{ \mathsf{B}(\alpha,\kappa+1) - \mathsf{B}(\alpha,p+\kappa+1) \right\} \\ &= \frac{1}{\mathsf{B}(\alpha,\kappa+1)} \sum_{j=1}^{p} \left\{ \mathsf{B}(\alpha,\kappa+j) - \mathsf{B}(\alpha,\kappa+j+1) \right\} \\ &= \frac{1}{\mathsf{B}(\alpha,\kappa+1)} \sum_{j=1}^{p} \mathsf{B}(\alpha+1,\kappa+j) = \sum_{j=1}^{p} \bar{\mathsf{B}}^{1,j-1}_{\alpha,\kappa+1}. \end{aligned} \tag{B.4.2}$$

Combining Equation (B.4.1) and Equation (B.4.2) we get the desired result. □

Proof of Proposition B.2.2. The proof is by induction. Let $\boldsymbol{\xi}_{j\cdot}$ be the j-th row of $\boldsymbol{\Xi}$. For $p = 1$, from a Poisson likelihood, we have

$$\mathsf{P}(\boldsymbol{\xi}_{1\cdot}) = \frac{1}{K_1^+!} \left(\gamma \bar{\mathsf{B}}^{1,0}_{\alpha,\kappa+1} \right)^{K_1^+} \mathrm{e}^{-\gamma \bar{\mathsf{B}}^{1,0}_{\alpha,\kappa+1}}$$

where K_1^+ is a number of nonzero elements of $\boldsymbol{\xi}_{1\cdot}$. It is same as Equation (B.2.3) with $p = 1$ and $K^+ = K_1^+$.

For $p \geq 2$, consider the conditional distribution of $\boldsymbol{\xi}_{p\cdot}$ given $\boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p-1\cdot}$, which is given by

$$\begin{aligned} \mathsf{P}(\boldsymbol{\xi}_{p\cdot} | \boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p-1\cdot}) = \mathrm{e}^{-\gamma \bar{\mathsf{B}}^{1,p-1}_{\alpha,\kappa+1}} \frac{(\gamma \bar{\mathsf{B}}^{1,p-1}_{\alpha,\kappa+1})^{K_p^{\mathrm{new}}}}{K_p^{\mathrm{new}}!} \\ \times \prod_{k \in J_p} \frac{m_{p-1,k} + \alpha}{p+\kappa+\alpha} \prod_{k \notin J_p} \frac{p - m_{p-1,k} + \kappa}{p+\kappa+\alpha}, \end{aligned}$$

where K_p^{new} is the number of new features sampled by the p-th customer and J_p is the set of dishes taken by the p-th customer, i.e., $J_p := \left\{ k \in [K_{p-1}^+] : \xi_{pk} = 1 \right\}$.

Let $K_p^+ := \sum_{j=1}^p K_j^{\text{new}} = K_{p-1}^+ + K_p^{\text{new}}$ and $K_1^{\text{new}} = K_1^+$. By the inductive hypothesis, we have

$$\begin{aligned}
\mathsf{P}(\boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p\cdot}) &= \mathsf{P}(\boldsymbol{\xi}_{p\cdot} | \boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p-1\cdot}) \mathsf{P}(\boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p-1\cdot}) \\
&= \mathrm{e}^{-\gamma \sum_{j=1}^p \bar{\mathsf{B}}_{\alpha,\kappa+1}^{0,j-1}} \frac{\gamma^{K_p^+}}{\prod_{j=1}^p K_j^{\text{new}}!} \prod_{k \in J_p} \frac{m_{p-1,k} + \alpha}{p + \kappa + \alpha} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p-1,k}, p-1-m_{p-1,k}} \\
&\quad \times \prod_{k \notin J_p} \frac{p - m_{p-1,k}}{p + \kappa + \alpha} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p-1,k}, p-1-m_{p-1,k}} \times \left(\bar{\mathsf{B}}_{\alpha,\kappa+1}^{1,p-1} \right)^{K_p^{\text{new}}}
\end{aligned}$$

Since

$$\begin{aligned}
\frac{m_{p-1,k} + \alpha}{p + \kappa + \alpha} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p-1,k}, p-1-m_{p-1,k}} &= \frac{m_{p-1,k} + \alpha}{p + \kappa + \alpha} \frac{\mathsf{B}(m_{p-1,k} + \alpha, p - m_{p-1,k} + \kappa)}{\mathsf{B}(\alpha, \kappa + 1)} \\
&= \frac{\mathsf{B}(m_{p-1,k} + 1 + \alpha, p - m_{p-1,k} + \kappa)}{\mathsf{B}(\alpha, \kappa + 1)} \\
&= \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p-1,k}+1, p-1-m_{p-1,k}}
\end{aligned}$$

and similarly

$$\frac{p - m_{p-1,k}}{p + \kappa + \alpha} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p-1,k}, p-1-m_{p-1,k}} = \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p-1,k}, p-m_{p-1,k}},$$

we have further that

$$\begin{aligned}
\mathsf{P}(\boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p\cdot}) &= \mathrm{e}^{-\gamma \sum_{j=1}^p \bar{\mathsf{B}}_{\alpha,\kappa+1}^{1,j-1}} \frac{\gamma^{K_p^+}}{\prod_{j=1}^p K_j^{\text{new}}!} \prod_{k \in J_p} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p-1,k}+1, p-1-m_{p-1,k}} \\
&\quad \times \prod_{k \notin J_p} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p-1,k}, p-m_{p-1,k}} \times \left(\bar{\mathsf{B}}_{\alpha,\kappa+1}^{1,p-1} \right)^{K_p^{\text{new}}} \\
&= \mathrm{e}^{-\gamma \sum_{j=1}^p \bar{\mathsf{B}}_{\alpha,\kappa+1}^{1,j-1}} \frac{\gamma^{K_p^+}}{\prod_{j=1}^p K_j^{\text{new}}!} \prod_{k \in J_p} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p,k}, p-m_{p,k}} \\
&\quad \times \prod_{k \notin J_p} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p,k}, p-m_{p,k}} \times \left(\bar{\mathsf{B}}_{\alpha,\kappa+1}^{1,p-1} \right)^{K_p^{\text{new}}} \\
&= \mathrm{e}^{-\gamma \sum_{j=1}^p \bar{\mathsf{B}}_{\alpha,\kappa+1}^{1,j-1}} \frac{\gamma^{K_p^+}}{\prod_{j=1}^p K_j^{\text{new}}!} \prod_{k=1}^{K_p^+} \bar{\mathsf{B}}_{\alpha,\kappa+1}^{m_{p,k}, p-m_{p,k}}
\end{aligned} \tag{B.4.3}$$

Note that $\prod_{j=1}^{p} K_j^{\text{new}}! / \prod_{\mathbf{u}\in\Delta_1} K_{\mathbf{u}}$ matrices generated by the above process have the same left-ordered form, hence $\mathsf{P}([\Xi])$ is obtained by multiplying $\mathsf{P}(\boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p\cdot})$ in Equation (B.4.3) by this quantity. □

Proof of Proposition B.2.3. By the well-known conjugacy result (Theorem 3.3 of Kim [49]),

$$\mu | \boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p\cdot} \sim \mathsf{CRM}(\Lambda_p, \{\omega_k^*, P_k\}_{k=1}^{K})$$

where

$$\Lambda_p(\mathrm{d}q, \mathrm{d}\omega) := \frac{\gamma}{\mathsf{B}(\alpha, \kappa+1)} q^{\alpha-1}(1-q)^{p+\kappa}\mathrm{d}q\Lambda_0(\mathrm{d}\omega),$$

ω_k^* are unique atoms that $\boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p\cdot}$ possess, and

$$\begin{aligned} P_k(\mathrm{d}q) &:= \frac{q^{m_{p,k}+\alpha-1}(1-q)^{p-m_{p,k}+\kappa}\mathrm{d}q}{\int_{(0,1]} q^{m_{p,k}+\alpha-1}(1-q)^{p-m_{p,k}+\kappa}\mathrm{d}q} \\ &= \frac{1}{\mathsf{B}(m_{p,k}+\alpha, p+1-m_{p,k}+\kappa)} q^{m_{p,k}+\alpha-1}(1-q)^{p-m_{p,k}+\kappa}\mathrm{d}q \end{aligned}$$

where $m_{p,k} := \sum_{j=1}^{p} \boldsymbol{\xi}_{j\cdot}(\omega_k^*)$.

For each atom ω_k^*, we have that

$$\begin{aligned} &\mathsf{P}\left(\boldsymbol{\xi}_{p+1\cdot}(\omega_k^*) = 1 | \boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{p\cdot}\right) \\ &= \frac{1}{\mathsf{B}(m_{p,k}+\alpha, p+1-m_{p,k}+\kappa)} \int_{(0,1]} q^{m_{p,k}+\alpha}(1-q)^{p-m_{p,k}+\kappa}\mathrm{d}q \\ &= \frac{1}{\mathsf{B}(m_{p,k}+\alpha, p+1-m_{p,k}+\kappa)} \mathsf{B}(m_{p,k}+1+\alpha, p+1-m_{p,k}+\kappa) \\ &= \frac{m_{p,k}+\alpha}{p+1+\kappa+\alpha}. \end{aligned} \tag{B.4.4}$$

This is equal to the mean of the Bernoulli distribution in (B.2.4).

For a small neighborhood $\mathrm{d}\omega$ around $\omega \in \Omega \setminus \{\omega_1^*, \ldots, \omega_K^*\}$, we have

$$\begin{aligned}
\mathsf{P}\left(\mathbf{z}_{p+1}(\mathrm{d}\omega) = 1 | \mathbf{z}_{1:p}\right) &= \mathsf{E}\left[\xi(\mathrm{d}\omega) | \mathbf{z}_{1:p}\right] \\
&= \frac{\gamma}{\mathsf{B}(\alpha, \kappa+1)} \int_{(0,1]} q^{\alpha}(1-q)^{p+\kappa} \mathrm{d}q \Lambda_0(\mathrm{d}\omega) \\
&= \gamma \frac{B(\alpha+1, p+\kappa+1)}{\mathsf{B}(\alpha, \kappa+1)} \Lambda_0(\mathrm{d}\omega) \\
&= \gamma \bar{\mathsf{B}}^{1,p}_{\alpha,\kappa+1} \Lambda_0(\mathrm{d}\omega)
\end{aligned} \tag{B.4.5}$$

This implies that on $\Omega \setminus \{\omega_1^*, \ldots, \omega_K^*\}$, $\boldsymbol{\xi}_{p+1\cdot}$ is a Poisson process with intensity measure $\gamma \bar{\mathsf{B}}^{1,p}_{\alpha,\kappa+1} \Lambda_0$, since $\boldsymbol{\xi}_{p+1\cdot}$ is completely random and Λ_0 is smooth. Thus, the number of new atoms in $\boldsymbol{\xi}_{p+1\cdot}$ follows Poisson distribution with rate $\gamma \bar{\mathsf{B}}^{1,p}_{\alpha,\kappa+1}$, which is equal to the rate parameter of the Poisson distribution in (B.2.5). □

B.4.2 Proofs of results in Section B.3.3

Proof of Lemma B.3.2. Let $(\theta_h)_{h\in[k]}$ be given. Then the random variable $S_k := |\mathsf{supp}_k(\mathbf{B})|$ is distributed as $\mathsf{Binom}(p, \pi_\theta)$ where the parameter π_θ satisfying $\pi_\theta \leq \sum_{h=1}^k \theta_h$. We define

$$\mathcal{E}_k := \left\{ (\theta_h)_{h\in\mathbb{N}} : \sum_{h=1}^k \theta_h \leq \frac{kt}{p^{3/2}}, \forall h \in [k] \right\}.$$

We bound $\Pi(S_k \geq t)$ as

$$\Pi(S_k \geq As) \leq \mathsf{E}\left[\Pi\left(S_k \geq t | (\theta_h)_{h\in[k]}\right) \mathbb{1}_{\mathcal{E}}\right] + \Pi(\mathcal{E}^c)$$

On the event $\mathcal{E}$, since $\pi_\theta \leq ktp^{3/2} \leq t/p$, and hence the tail bound of the binomail distribution in (3.7.1) in Chapter 3 implies that

$$\begin{aligned}
\Pi(S_k \geq t|(\theta_h)_{h\in[k]}) &= \Pi\left(S_k \geq \frac{t}{p}p\middle|(\theta_h)_{h\in[k]}\right) \\
&\leq \left[\left\{\pi_\theta \frac{p}{t}\right\}^{t/p} \mathrm{e}^{t/p}\right]^p \\
&\leq \left\{\frac{k}{p^{1/2}}\right\}^t \mathrm{e}^t \\
&\leq \left\{\frac{1}{p^{1/6}}\right\}^t \mathrm{e}^t = \mathrm{e}^{t-\frac{t}{6}\log p}
\end{aligned}$$

on the event $\mathcal{E}_k$.

Using the union bound, we have

$$\begin{aligned}
\Pi(\mathcal{E}_k^c) &= \Pi\left(\bigcup_{h=1}^k \{\theta_h : \theta_h \geq t/p^{3/2}\}\right) \\
&\leq k\Pi\left(\theta_1 > t/p^{3/2}\right) \\
&= \frac{k}{\mathrm{B}(1, p^2+1)} \int_{t/p^{3/2}}^1 (1-\theta)^{p^2} \mathrm{d}\theta \\
&\leq k\left(1 - \frac{t}{p^{3/2}}\right)^{p^2+1} \\
&\leq p^{1/3} \exp\left(-p^{1/2}/2\right),
\end{aligned}$$

which completes the proof □

Proof of Lemma B.3.3. We start with observing

$$\begin{aligned}
\Pi(\|\mathbf{B} - \mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n) &\geq \Pi(\|\mathbf{B} - \mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n | K = k_{0n})\Pi(K = k_{0n}) \\
&\gtrsim \gamma_n^{k_{0n}} \mathrm{e}^{-k_{0n}\log k_{0n}} \Pi(\|\mathbf{B} - \mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n | K = k_{0n}).
\end{aligned}$$

By Lemma 3.7.3, we have

$$
\begin{aligned}
&\Pi\left(\|\mathbf{B}-\mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n | (\theta_k)_{k\in[k_{0n}]}, K=k_{0n}\right) \\
&\geq \prod_{k=1}^{k_{0n}} \mathsf{P}\left(\|\boldsymbol{\beta}_k-\boldsymbol{\beta}_{0n,k}\|_2 \leq \frac{\eta_n}{\sqrt{k_{0n}}} \middle| \theta_k, K=k_{0n}\right) \\
&\geq \prod_{k=1}^{k_{0n}} \theta_k^{s_n}(1-\theta_k)^{p_n-s_n} \exp\left(-\|\boldsymbol{\beta}_{0n,k}\|_1 - \frac{\eta_n}{\sqrt{2k_{0n}}}\right)\left(\frac{\eta_n}{\sqrt{2s_nk_{0n}}}\right)^{s_n} \\
&\geq \exp\left(-\|\mathbf{B}_{0n}\|_1 - \frac{\eta_n\sqrt{k_{0n}}}{\sqrt{2}}\right)\left(\frac{\eta_n}{\sqrt{2s_nk_{0n}}}\right)^{s_nk_{0n}} \prod_{k=1}^{k_{0n}} \theta_k^{s_n}(1-\theta_k)^{p_n-s_n} \\
&\geq \exp\left(-\|\mathbf{B}_{0n}\|_1 - C_1 s_n k_{0n}\log(p_n \vee \eta_n^{-1})\right) \prod_{k=1}^{k_{0n}} \theta_k^{s_n}(1-\theta_k)^{p_n-s_n}
\end{aligned}
$$

for some universal constant $C_1 > 0$. Since $\mathsf{B}(1, p_n^2+1) < \mathsf{B}(1,1) = 1$ and $p_n^{1+\delta} > p_n - s_n$, it follows that

$$
\begin{aligned}
&\mathsf{E}\left\{\theta_1^{s_n}(1-\theta_1)^{(p_n-s_n)}\right\} \\
&= \frac{1}{\mathsf{B}(1,p_n^2+1)} \int_0^1 \theta^{s_n}(1-\theta)^{p_n^2+p_n-s_n}\mathrm{d}\theta \\
&\geq \int_0^{p_n^{-2}} \theta^{s_n}(1-\theta)^{2p_n^2}\mathrm{d}\theta \\
&\geq \int_0^{p_n^{-2}} \theta^{s_n}\mathrm{d}\theta \left(1-\frac{1}{p_n^2}\right)^{2p_n^2} \\
&\geq \frac{1}{s_n+1}\left(\frac{1}{p_n^2}\right)^{s_n} \mathrm{e}^{-4} \\
&\geq \mathrm{e}^{-C_2 s_n \log p_n},
\end{aligned}
$$

for some $C_2 > 0$, where the third inequality is due to the inequality $(1-x)^{1/x} \geq \mathrm{e}^{-2}$ for $0 < x < 1/2$. Since $\theta_1, \ldots, \theta_K$ are independent given K, we

have that

$$\begin{aligned}
&\Pi(\|\mathbf{B}-\mathbf{B}_{0n}\|_{\mathrm{F}} \leq \eta_n | K = k_{0n}) \\
&\geq \mathrm{e}^{-\|\mathbf{B}_{0n}\|_1 - C_1 s_n k_{0n} \log p_n} \mathsf{E}\left[\prod_{k=1}^{k_{0n}} \theta_k^{s_n} (1-\theta_k)^{p_n - s_n}\right] \\
&\gtrsim \mathrm{e}^{-\|\mathbf{B}_{0n}\|_1 - C_1 s_n k_{0n} \log p_n} \mathrm{e}^{-C_2 s_n k_{0n} \log p_n}.
\end{aligned}$$

By (A2), we have $\|\mathbf{B}_{0n}\|_1 \leq \sqrt{s_n} k_{0n} \|\mathbf{B}_{0n}\| \lesssim s_n k_{0n}$, which completes the proof. □

B.4.3 Proof of Theorem B.3.5

Proof of Theorem B.3.5. To simplify notation, we write $\mathsf{P}_0 := \mathsf{P}_{\mathbf{\Sigma}_{0n}}$ and $\mathsf{E}_0 := \mathsf{E}_{\mathbf{\Sigma}_{0n}}$.

Let $\gamma_n := p_n^{-As_n^2}$. Fix $\mathbf{\Sigma}_{0n} \in \mathcal{C}_{0n}$. By Lemma B.3.4, we have that there is a Borel measurable set $\mathfrak{A}_n \in \sigma(\mathbf{Y}_{1:n})$ with $\mathsf{P}_0(\mathfrak{A}_n) \lesssim 1/\log n$, on which

$$D_n := \int \prod_{i=1}^{n} \frac{f_{\mathbf{\Sigma}}(\mathbf{Y}_i)}{f_{\mathbf{\Sigma}_{0n}}(\mathbf{Y}_i)} \mathrm{d}\Pi_n(\mathbf{\Sigma}) \geq \gamma_n^{k_{0n}} \mathrm{e}^{-C_1 s_n k_{0n} \log p_n}. \tag{B.4.6}$$

for some universal constant $C_1 > 0$.

Define

$$\mathcal{V}_n := \left\{\mathbf{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2 \mathbf{I} : \mathsf{K}^+(\mathbf{B}) > k_{0n}\right\}.$$

By Lemma B.3.1, we have

$$\begin{aligned}
\mathsf{E}_0[\Pi_n(\mathcal{V}_n | \mathbf{Y}_{1:n})] &\leq \mathsf{E}_0\left[\frac{\Pi_n(\mathsf{K}^+(\mathbf{B}) > k_{0n})}{D_n} \mathbb{1}_{\mathfrak{A}_n}\right] + \mathsf{P}_0(\mathfrak{A}_n^c) \\
&\lesssim \gamma_n \mathrm{e}^{C_1 s_n k_{0n} \log p_n} + \frac{1}{\log n}.
\end{aligned}$$

Thus for sufficiently large $A > 0$ such that $A > C_1$, it follows that $\mathsf{E}_0[\Pi_n(\mathsf{K}^+(\mathbf{B}) > k_{0n} | \mathbf{Y}_{1:n})]$ converges to zero as $n \to \infty$.

For the posterior contraction of the covariance matrix, we construct a sieve $\mathcal{F}_n$ as

$$\mathcal{F}_n := \left\{ \boldsymbol{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2 \mathbf{I} : |\mathsf{supp}_{k_{0n}}(\mathbf{B})| \leq C_2 s_n^2 k_{0n} \right\},$$

for $C_2 > 0$ which is to be specified later. Then by the assumption (A4) that implies $k_{0n} \lesssim p_n^{1/3}$, we have $\Pi(\mathcal{F}_n^c) \leq \exp(-C_2 s_n^2 k_{0n} \log p_n)$. Thus $\mathsf{E}_0[\Pi_n(\mathcal{F}_n^c | \mathbf{Y}_{1:n})]$ converges to zero if $C_2 > A + C_1$. Let

$$\mathcal{U}_n^\star := \left\{ \boldsymbol{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2 \mathbf{I} : \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\| \geq M\epsilon_n, |\mathsf{supp}_{k_{0n}}(\mathbf{B})| \leq C_2 s_n^2 k_{0n} \right\}.$$

with

$$\epsilon_n := c_n \sqrt{\frac{s_n^2 k_{0n} \log p_n}{n}}.$$

Then Lemma 3.7.5 in Chapter 3 implies that there is a test function ϕ_n such that

$$\begin{aligned}
\mathsf{E}_0 \left[\Pi_n \left(\mathcal{U}_n^\star | \mathbf{Y}_{1:n} \right) \right] &\leq \mathsf{E}_0 \phi_n + \mathsf{E}_0 \left[(1 - \phi_n) \Pi_n \left(\mathcal{U}_n^\star | \mathbf{Y}_{1:n} \right) \mathbb{1}_{\mathfrak{A}_n} \right] + \mathsf{P}_0(\mathfrak{A}_n^c) \\
&\lesssim \mathsf{E}_0 \phi_n + \mathrm{e}^{(A+C_1) s_n^2 k_{0n} \log p_n} \sup_{\boldsymbol{\Sigma} \in \mathcal{U}_n^\star} \mathsf{E}_{\boldsymbol{\Sigma}} (1 - \phi_n) + \frac{1}{\log n} \\
&\lesssim \mathrm{e}^{C_3 s_n^2 k_{0n} \log p_n - C_4 M^{1/2} n \epsilon_n^2 / c_n^2} \\
&\quad + \mathrm{e}^{(A+C_1) s_n^2 k_{0n} \log p_n + C_5 (s_n^2 k_{0n} + s_n) - C_6 M n \epsilon_n^2} + \frac{1}{\log n} \\
&\lesssim \exp\left((C_7 - C_8 M^{1/2}) s_n^2 k_{0n} \log p_n \right) + \frac{1}{\log n}
\end{aligned}$$

for some universal positive constants $C_3, \ldots, C_8$. Hence for sufficiently large $M > 0$, it follows that

$$\mathsf{E}_0 \left[\Pi_n \left(\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\| > M\epsilon_n | \mathbf{Y}_{1:n} \right) \right] = o(1).$$

For the posterior consistency of the factor dimensionality, we further assume that $\boldsymbol{\Sigma}_{0n}$ satisfies the condition (A5), i.e., $\boldsymbol{\Sigma}_{0n} \in \mathcal{C}_{0n}^\star$. We have proven that $\mathsf{E}_0[\Pi_n(\mathsf{K}^+(\mathbf{B}) > k_{0n} | \mathbf{Y}_{1:n})] = o(1)$. For the event $\{\mathsf{K}^+(\mathbf{B}) < k_{0n}\}$, we invoke similar arguments used in the proof of Theorem 3.3.2 in Chapter 3,

which states that

$$\mathsf{E}_0\left[\Pi_n\left(\mathsf{K}^+(\mathbf{B}) < k_{0n}|\mathbf{Y}_{1:n}\right)\right] \leq \mathsf{E}_0\left[\Pi_n\left(\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| > M\epsilon_n|\mathbf{Y}_{1:n}\right)\right] = o(1).$$

This completes the proof. □

Bibliography

[1] Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

[2] Anthony, M. and Bartlett, P. L. (2001). *Neural network learning: Theoretical foundations*. Cambridge university press.

[3] Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633.

[4] Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

[5] Bai, J. and Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, 25(1):52–60.

[6] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

[7] Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285.

[8] Bergstra, J., Desjardins, G., Lamblin, P., and Bengio, Y. (2009). Quadratic polynomials learn better image features. *Technical report, Technical Report 1337, Département d'Informatique et de Recherche Operationnelle*.

[9] Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics*, 120(1):387–422.

[10] Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306.

[11] Blanchard, G., Bousquet, O., and Massart, P. (2008). Statistical performance of support vector machines. *The Annals of Statistics*, 36(2):489–531.

[12] Bunea, F., Giraud, C., Luo, X., Royer, M., and Verzelen, N. (2020). Model-assisted variable clustering: minimax-optimal recovery and algorithms. *The Annals of Statistics*, Accepted.

[13] Carlile, B., Delamarter, G., Kinney, P., Marti, A., and Whitney, B. (2017). Improving deep learning by inverse square root linear units (ISRLUs). *arXiv preprint arXiv:1710.09967*.

[14] Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*, pages 2051–2059.

[15] Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456.

[16] Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101.

[17] Chen, B., Chen, M., Paisley, J., Zaas, A., Woods, C., Ginsburg, G. S., Hero, A., Lucas, J., Dunson, D., and Carin, L. (2010). Bayesian inference of the number of factors in gene-expression analysis: application to human virus challenge studies. *BMC bioinformatics*, 11(1):552.

[18] Chui, C. K. and Li, X. (1992). Approximation by ridge functions and neural networks with one hidden layer. *Journal of Approximation Theory*, 70(2):131–141.

[19] Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*.

[20] Costarelli, D. and Sambucini, A. R. (2018). Approximation results in orlicz spaces for sequences of kantorovich max-product neural network operators. *Results in Mathematics*, 73(1):15.

[21] Costarelli, D. and Spigler, R. (2018). Solving numerically nonlinear systems of balance laws by multivariate sigmoidal functions approximation. *Computational and Applied Mathematics*, 37(1):99–133.

[22] Costarelli, D. and Vinti, G. (2017). Saturation classes for max-product neural network operators activated by sigmoidal functions. *Results in Mathematics*, 72(3):1555–1569.

[23] Costarelli, D. and Vinti, G. (2018). Estimates for the neural network operators of the max-product type with continuous and p-integrable functions. *Results in Mathematics*, 73(1):12.

[24] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

[25] Eldan, R. and Shamir, O. (2016). The power of depth for feedforward neural networks. In *Conference on Learning Theory*, pages 907–940.

[26] Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

[27] Fan, J., Han, X., and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035.

[28] Fan, J., Ke, Y., Sun, Q., and Zhou, W.-X. (2020). FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. *Journal of the American Statistical Association*, Accepted.

[29] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

[30] Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320.

[31] Fan, J., Liu, H., and Wang, W. (2018). Large covariance estimation through elliptical factor models. *The Annals of Statistics*, 46(4):1383.

[32] Fan, J., Xue, L., and Yao, J. (2017). Sufficient forecasting using factor models. *Journal of Econometrics*, 201(2):292–306.

[33] Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2003). Do financial variables help forecasting inflation and real activity in the Euro area? *Journal of Monetary Economics*, 50(6):1243–1255.

[34] Frankle, J. and Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

[35] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407.

[36] Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192.

[37] Gao, C. and Zhou, H. H. (2015). Rate-optimal posterior contraction for sparse PCA. *The Annals of Statistics*, 43(2):785–818.

[38] Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in neural information processing systems*, pages 475–482.

[39] Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 315–323.

[40] Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

[41] Hagerup, T. and Rüb, C. (1990). A guided tour of chernoff bounds. *Information processing letters*, 33(6):305–308.

[42] Han, Q. (2017). Bayes model selection. *arXiv preprint arXiv:1704.07513*.

[43] Han, S., Pool, J., Tran, J., and Dally, W. (2015). Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, pages 1135–1143.

[44] Hochreiter, S., Clevert, D.-A., and Obermayer, K. (2006). A new summarization method for affymetrix probe level data. *Bioinformatics*, 22(8):943–949.

[45] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

[46] Imaizumi, M. and Fukumizu, K. (2019). Deep neural networks learn non-smooth functions effectively. In *International Conference on Artificial Intelligence and Statistics*, pages 869–878.

[47] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456.

[48] Kim, D., Choi, Y., and Kim, Y. (2019). Understanding and improving virtual adversarial training. *arXiv preprint arXiv:1909.06737*.

[49] Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *The Annals of Statistics*, pages 562–588.

[50] Kim, Y., Choi, H., and Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673.

[51] Kim, Y., Ohn, I., and Kim, D. (2018). Fast convergence rates of deep neural networks for classification. *arXiv preprint arXiv:1812.03599*.

[52] Klimek, M. D. and Perelstein, M. (2018). Neural network-based approach to phase space integration. *arXiv preprint arXiv:1810.11509*.

[53] Kneip, A. and Sarda, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics*, 39(5):2410–2447.

[54] Knowles, D. and Ghahramani, Z. (2011). Nonparametric bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, pages 1534–1552.

[55] Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

[56] Latala, R. (2005). Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282.

[57] Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723.

[58] Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867.

[59] Li, B., Tang, S., and Yu, H. (2019). Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. *arXiv preprint arXiv:1903.05858*.

[60] Lin, Y. (2004). A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82.

[61] Liu, B., Wang, M., Foroosh, H., Tappen, M., and Pensky, M. (2015). Sparse convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 806–814.

[62] Louizos, C., Welling, M., and Kingma, D. P. (2018). Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*.

[63] Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829.

[64] Martin, R., Mess, R., and Walker, S. G. (2017). Empirical bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847.

[65] McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

[66] Meeds, E., Ghahramani, Z., Neal, R. M., and Roweis, S. T. (2007). Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems*, pages 977–984.

[67] Mhaskar, H. N. (1993). Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, 1(1):61–80.

[68] Miller, K., Jordan, M. I., and Griffiths, T. L. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, pages 1276–1284.

[69] Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2924–2932.

[70] Navarro, D. J. and Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric bayesian approach. *Neural computation*, 20(11):2597–2628.

[71] Ohn, I. and Kim, Y. (2019). Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21(7):627.

[72] Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

[73] Paisley, J. and Carin, L. (2009). Nonparametric factor analysis with beta process priors. In *International Conference on Machine Learning*, pages 777–784.

[74] Park, C. (2009). Convergence rates of generalization errors for margin-based classification. *Journal of Statistical Planning and Inference*, 139(8):2543–2551.

[75] Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42(3):1102–1130.

[76] Petersen, P. and Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330.

[77] Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Dickstein, J. S. (2017). On the expressive power of deep neural networks. In *International Conference on Machine Learning*, pages 2847–2854.

[78] Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

[79] Rockova, V. and George, E. I. (2016). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

[80] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, Accepted.

[81] Shen, X., Wong, W. H., et al. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, 22(2):580–615.

[82] Silva, A. P. D. (2011). Two-group classification with high-dimensional correlated data: A factor model approach. *Computational Statistics & Data Analysis*, 55(11):2975–2990.

[83] Srivastava, S., Engelhardt, B. E., and Dunson, D. B. (2017). Expandable factor analysis. *Biometrika*, 104(3):649–663.

[84] Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

[85] Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607.

[86] Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

[87] Suzuki, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations*.

[88] Tarigan, B. and Van De Geer, S. A. a. (2006). Classifiers of support vector machine type with ℓ_1 complexity regularization. *Bernoulli*, 12(6):1045–1076.

[89] Teh, Y. W. and Gorur, D. (2009). Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, pages 1838–1846.

[90] Teh, Y. W., Grür, D., and Ghahramani, Z. (2007). Stick-breaking construction for the indian buffet process. In *Artificial Intelligence and Statistics*, pages 556–563.

[91] Telgarsky, M. (2017). Neural networks and rational functions. In *International Conference on Machine Learning*, pages 3387–3393.

[92] Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the indian buffet process. In *Artificial Intelligence and Statistics*, pages 564–571.

[93] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166.

[94] Tsybakov, A. B. and van de Geer, S. A. (2005). Square root penalty: adaptation to the margin in classification and in edge estimation. *The Annals of Statistics*, 33(3):1203–1224.

[95] van de Geer, S. A. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

[96] Van Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756.

[97] Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. (2016). Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2074–2082.

[98] Wuraola, A. and Patel, N. (2018). SQNL: A new computationally efficient activation function. In *International Joint Conference on Neural Networks*, pages 1–7.

[99] Xie, F., Xu, Y., Priebe, C. E., and Cape, J. (2018). Bayesian estimation of sparse spiked covariance matrices in high dimensions. *arXiv preprint arXiv:1808.07433*.

[100] Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114.

[101] Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation*, 15(4):915–936.

[102] Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

[103] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85.

[104] Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(Mar):1081–1107.

국문초록

본 논문은 세 가지 기계 학습 알고리즘의 점근 성질을 연구한다. 처음 두 장은 지도학습에 사용되는 깊은 신경망 학습을 다루며 마지막 장은 고차원 요인 모형의 베이지안 추정 방법에 대하여 연구한다.

첫 번째 장에서는 깊은 신경망 분류기에 대하여 연구한다. 우리는 힌지 손실 함수로 학습한 깊은 신경망 분류기가 몇 가지 확률 모형에 대해 빠른 수렴 속도를 달성함을 보였다.

두 번째 장에서는 깊은 신경망의 희소 학습에 대하여 연구한다. 우리는 경험 위험과 희소성을 부여하는 벌점 함수를 더한 목적 함수를 최소화하는 학습 방법을 제안하였다. 우리는 제안하는 깊은 희소 신경망 추정량에 대한 신의 부등식을 얻었으며, 이를 통해 몇 가지 통계 문제에서의 수렴 속도를 구하였다. 특히 제안하는 깊은 희소 신경망 추정량은 비모수 회귀 문제에서 적응적으로 최소최대 최적성을 달성함을 보였다.

마지막 장은 고차원 요인 모형에서 베이지안 학습의 점근 성질을 연구한다. 우리는 모수가 두개인 인도부페과정을 기반으로 한 사전분포를 제안하였다. 제안한 사전분포로부터 유도된 사후분포가 공분산 행렬을 거의 최적의 수렴 속도로 추정함과 동시에 요인 차원을 일관되게 추정할 수 있음을 증명하였다.

주요어: 비모수 베이지안, 깊은 신경망, 빠른 수렴속도, 요인모형, 최소최대 최적성, 사후 수렴속도, 희소성

학　번: 2014–21216